Optimizing Grouping Data from the Open Ended Question about the Community Development Needs of the People

Assistant Professor Dr. Prapai Sridama

***Abstract*—**The objectives in this research are the data classification by the open ended question by support vector machine (OLE_SVM) model and the data clustering by K-mean algorithm. Many procedures is used to decrease the content of 1,500 the open ended questions as follows: the word segmentation, the feature extraction, the stopword and stemming, the indexing and the feature selection. This research collects from 1,500 people in Laksi district at Bangkok about the demand for community development. The efficiency of data classification of the OEQ_SVM model is compared with ID3 model. It can be seen that the OEQ_SVM model gives the data classification values higher than the ID3 model in any groups of data sets. The highest average value of the OEQ_SVM model is equal 94.2 percent while the highest average value of the ID3 model is equal 79.3 percent. This research uses eight groups for using in both models. However, the highest need of people is the security of life and demand. Then this research selects this group for the data clustering. The result of the data clustering can divide the public needs and demand. The greatest needs and demand is equal 259 number or 29.53 percent about the providing for installation of CCTV.

***Keywords*—**ID3 algorithm, SVM algrothm, K-mean algorithm, Decision tree.

I. Introduction

BANGKOK is the center of government, transportation, communications, trade and investment, transportation, financial, education and more administratively. Bangkok is divided into 50 districts in the present. In each district is divided into 10 departments as follows: administration, registration, environment and sanitation, civil construction, revenue, cleanliness and parks, education, municipal, community and social welfare, and finance. This research focuses on department of community and social welfare because this department is closer to the people as possible and to touch the lives of people directly. In addition, this department is exposure to the problems faced by the people in each country. Moreover, this department is also legislation or policies to point as possible.

The department of community and social welfare is responsible for the implementation of community development and social welfare, both physical, economic, social, health and quality of life, which are promoting the participation of citizens, to develop environment and housing, community update, promoting career, and others. The Department of Community Development and Social Services is responsible for 1) the election of community board 2) the establishment of a community by Bangkok with the community and the community 3) to make identification cards and 4) to seek universal health insurance in Bangkok. However, community problems and society has not healed as much as it should after applying good governance principles used in the management of Bangkok. In addition, not all problems have been resolved rightly. Then, the involvement of the public administration sector is important to contribute to the answer to the problem.

From a focus on local development, the researcher interests the demand for community development. The people should be involved in shaping the solutions in the community. The demands for community development by open-ended questions are used to collect from people. This research collects from 1,500 people of Laksi district. The cause of selection in this district is cultural diversity and career.

The open-ended questions are classified by learning with computer and to apply natural language processing technique. The natural language processing has two characters as follows: 1) data clustering and 2) classification or categorization. The data clustering can divide by type of document that the data clustering does not know how many groups. The data classification divides groups by the content of the document. This technique knows number of groups before dividing the document.

The classification of documents focuses on English document in the present but Thai document lacks the classification. The objective in this research are development of Thai document classification and survey the needs of the development community. Theories related research are word segmentation, stop-word list removal, stemming, feature extraction, indexing, decision tree, and support vector machine.

This paper is divided four main sections after this section as follows. The theories related research are shown in first section after that the open ended question by support vector machine (OEQ_SVM) model is presented in second section and then the results of the OEQ_SVM model are presented. Lastly, the conclusions of this research are proposed.

Assistant Professor Dr. Prapai Sridama is with Department of Computer Science, Faculty of Science and Technology, Bansomdejchaopraya Rajhaphat University, Bangkok, Thailand.

II. Theories Related Research

A. Word segmentation

The important problem of Thai word segmentation is the written word to each other and no punctuation. In contrast, English word uses space for dividing words. Many researchers create to develop word segmentation techniques such as rule base approach, algorithm approach, dictionary approach and corpusbase approach. Each approach gives different accuracy value and speed of processing. The studying about Thai word segmentation founds that the scope of word is big problem because Thai language dose not divide in each word or a sentence. In addition, Thai language dose not has fixed rules on the use spaces the written language. Furthermore, Thai language has loanwords and transliterated words. From education about comparing the powerful words founds that, the efficiency technique of word segmentation is the longest matching technique. This technique checks string from left to right and compares with Thai dictionary. If the check is found to have more than one syllable then this technique selects longest syllable. After that, this technique checks all of the strings. If the longest syllable does not found the word of Thai dictionary then back tracking technique is used [1], [2].

B. Stop-word list removal

The stop-word list removal is to bring an insignificant issue. After to make stop-word list removal, The sentence does not has to make significant changes. The advantage of stop-word list removal is to save the memory and the time of processing. The stop-words are preposition, conjunction and pronoun. Many documents are found many stop-words that these words are useless for the classification. However, the stop-word list removal should make before make the indexing [3].

C. Finding roots or stemming

The finding root is to search the synonymous words. These synonymous words can be combined as a single unit. This step should make before make the indexing. It can decrease the indexing and can increase the efficiency of the classification too. However, the finding root uses expert human in Thai language to define these words [3].

D. Feature extraction

The objective in this procedure is feature extraction of document. This procedure extracts features of documents and decreases the size of the sentences. It defines words for replacing of the features of documents. The popular in feature extraction is the bag of words, which are vector types. The element of vectors can use Boolean or word frequency [4], [5].

E. Indexing

The computer can not directly classify document, which is natural language. Then document is converted to a type that the computer can learn from this type. The converted document is called the indexing. The indexing creates document representation for learning procedure. The objective of the indexing is to calculate values, which are used attribute values of document. Some time, this procedure is called the term weighting. The first step of indexing is making vectors of document representation. After that, it creates metric of groups from all vectors [6]. This research uses the frequency of words, which appear in open-ended questions. These words are processed in the word segmentation procedure. If words appear at many open-ended questions then the frequency values are high values. The relation between words (Words: w) and all open-ended questions (o) by 2 dimension vector. The indexing is presented at figure 1.

	W_1	W_2	...	W_i	...	W_j
O_1	W_{11}	W_{12}	...	W_{1i}	...	W_{1v}
O_2	W_{21}	W_{22}	...	W_{2i}	...	W_{2v}
...	...	...	...	...	...	...
O_k	W_{k1}	W_{k2}	...	W_{ki}	...	W_{kv}

Fig. 1: Vector space model

F. Feature selection

Many learning algorithms for document classification cannot support the a lot of number of features then decreasing contents of documents are important procedure. The information gain (IG) is used to decrease the feature of document. The IG calculates from the number of bits, which are used for group prediction. Let $S_i, \ldots, S_k$ is a set, which is the set of possible group. The IG of word (w) [7], [8].

$$IG(w) = \sum_{i=1}^{k} P(s_i)\log P(s_i) + P(w)\sum_{i=1}^{k} P(s_i \mid w)\log P(s_i \mid w) + P(\overline{w})\sum_{i=1}^{k} P(s_i \mid \overline{w})\log P(s_i \mid \overline{w}), \quad (1)$$

Where $P(s_i)$ is calculated by the fraction of the number of open-ended questions, which are in s_i with the number of all open-ended questions. $P(w)$ is calculated by the fraction of the number of open-ended questions, which have word (w) with the number of all open-ended questions. $P(s_i \mid w)$ is calculated by the fraction of word, which is in s_i and has word (w) with all open-ended questions. $P(s_i \mid \overline{w})$ is calculated by the fraction of word, which is in s_i and has not word (w) with all open-ended questions.

After to calculate IG value in each the feature, all IG values are sorted descending. If IG values are less than satisfactory then these IG values are removed. [9].

G. Classifier algorithm

The supervised learning can divide two procedure that are 1) learning to create the group the document model and 2) classification of documents of interest. The classification of documents of interest verifies the similar to the document model.

- Decision tree: the decision tree includes nodes and bottom node is represented from the category. To create branches consider from real values of attributes. These attribute values are calculate by information gain[10]. This research uses ID3 algorithm for creating the decision tree.

Support vector machine (SVM): the concept of SVM is used to find the decision by the plane is that SVM is divided into two parts [10]. The SVM uses linear equation for divided two

groups. The straight line should be in the middle between two groups. The distance from the boundary is between the two groups as much as possible. The SVM uses map function for removed data from input spate to feature space and created kernel function on feature space. Let (x_i, y_i), …, (x_n, y_n) are learning sets, n is the number of sampling, m is the number of dimensions, and y is the result values, which are between +1 and -1. This equation is present at (2).

$$(x_i, y_i), \ldots, (x_n, y_n) \text{ when } x \in R^m, y \in \{+1, -1\} \quad (2)$$

The linear problem of data dimension is divided two groups by (3).

$$(w^* x)b = 0, \quad (3)$$

where w is weight value and b is bias. The equation for data classification is shown at (4).

$$(w^* x)b > 0 \text{ if } y_i = +1 \text{ and} \quad (4)$$

$$(w^* x)b < 0 \text{ if } y_i = -1$$

However, this research uses linear kernel function and defines a parameter (P), which is equal 1.

III. The Open Ended Question by Support Vector Machine (OEQ_SVM) model

The research decreases contents of 1,500 open-ended questions (4,500 items) that many procedures are called pre-processing. After these procedures are processed then this research can get eight problem groups. The problem groups are drugs, maladies, security of life and property, education, family, psychological and moral decadence, health, and disadvantaged groups. All procedures of OEQ_SVM model in this research are shown at figure 2.

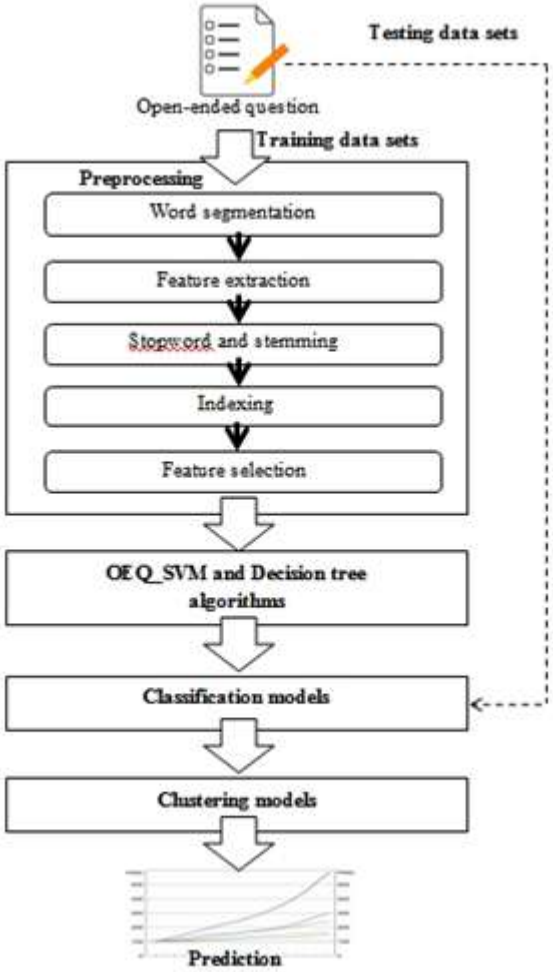

Fig. 2: Procedures of OEQ_SVM model

From figure 2, the preprocessing procedure is created by theories, which are explained in theories related research section. The decreased content of open-ended question is gotten by this procedure. Next step is OEQ_SVM and decision tree algorithms. This step is used to develop both OEQ_SVM and decision tree algorithms. Equations of (2) to (4) are used for the OEQ_SVM model and ID3 algorithm is use to create decision tree model.

Next step is to compare the efficiency of OEQ_SVM model and decision tree model. The recall equation, precision equation and F-measure are used to benchmarks.

$$recall = \frac{p}{p+r}, \quad (5)$$

$$percision = \frac{p}{p+q}, \quad (6)$$

$$F - measure = \frac{2*(\text{precision}*\text{recall})}{\text{precision} + \text{recall}}, \quad (7)$$

where p is the number of open-ended items, which are in C_i and to predict that p is in C_i.

r is the number of open-ended items, which are in C_i and to predict that r is not in C_i.

q is the number of open-ended items, which are not in C_i and to predict that q is not in C_i.

C_i is the open-ended type, which is attended.

After past classification model, the data classification of community issues is received. After that, the maximum number of the data classification is used to clustering data. The K-means algorithm is used to find demand for most of the group selected.

IV. The Results of the OEQ_SVM Model

The results of data classification of community issues are presented in this section. This research compares the efficiency models between the OEQ_SVM model and Decision Tree. The problem is divided into 10 categories as follows: drugs, maladies, security of life and property, issues in education, family issues, psychological and moral decadence, health problems, and problems disadvantage groups.

TABLE I: Comparing of data classification by F-Measure

Number of testing set	ID3	OEQ_SVM
4500	0.621	0.938
4000	0.744	0.921
3500	0.737	**0.942**
3000	0.742	0.897
2500	**0.793**	0.904
2000	0.724	0.924
1500	0.753	0.885
1000	0.731	0.846
500	0.722	0.873
100	0.687	0.837
Average	**0.725**	**0.897**

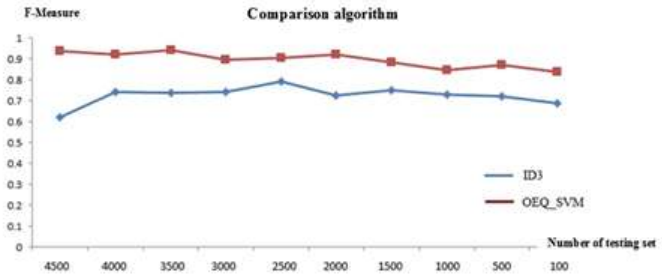

Fig. 3: Comparison between OEQ_SVM algorithm and ID3 algorithm

The results from this research, its can show that the highest average value of OEQ_SVM is 0.942 (94.2%) at 3,500 testing sets. And the highest average value of ID3 is 0.793 (79.3%) at 2,500 testing sets. In other words, the maximum efficiency of the OEQ_SVM algorithm is 3,500 testing sets and the maximum efficiency of the ID3 algorithm is 2,500 testing sets.

The security of life and property group is highest selected. It is 877 items. Then this group is used to divide by the clustering algorithm. However, this research uses the K-Mean algorithm.

TABLE II: The Public Needs and Demand

Groups by K-mean algorithm	The number of public needs and demand	Average
Providing for the installation of CCTV	259	29.53
Providing additional spot lighting	204	23.26
The guard house	159	18.13
Set up checkpoints and raids by policeman	194	22.12
other	**61**	6.96
	877	**100**

The table 2 presents about the public needs and demand, which are created by K-mean algorithm. It can be show that the providing for the installation of CCTV is greatest demanded that it is equal 259 needs or 29.53 percent. The second is to increase the lighting that the demand is equal 204 needs or 23.29 percent.

V. Conclusion

This research focuses on the data classification by the open ended question by support vector machine (OEQ_SVM) model and data clustering by K-means algorithm. The open ended question is used to survey at Laksi district at Bangkok. The number of people is equal 1,5000 persons for answering. Many procedures are used to prepare training sets and data sets as follows: the word segmentation, the feature extraction, stopword and stemming, the indexing, and feature selection. This research compares the efficiency of the data classification between the OEQ_SVM model and ID3 model. It can be seen that the OEQ_SVM model classified information than the ID3 models in each segment. The OEQ_SVM model can be best classified information for the 3,500 data sets while the ID3 model can be best classified information for the 2,500 data sets. However, the OEQ_SVM model gives the greatest average value of the F-measure at 94.2 percent while the ID3 model offers the greatest average value of the F-measure at 79.3 percent. In addition, the result of data clustering procedure for finding the highest needs from 1,500 persons is 29.53 percent that they want to install CCTV.

References

[1] P. Charoenpornsawat, "Feature-based Thai word segmentation," M.S. thesis, Dept. Computer Engineering, Chulalongkorn University, Bangkok, Thailand, 1999.

[2] J. Markpong, T. Issara, W. Chai, and F. Sadaoki, "Lexical units for Thai LVCSR," *Original research article speech communication*, vol. 51, pp. 379-389, 2009.

[3] C. Jaruskulchai, "An automatic indexing for Thai text retrieval," PhD thesis, George Washington University, USA, 1998.

[4] Aas., Eikvil, "Text categorization: a survey," Report Nurwegian computing center, 1999.

[5] N. Pirconsup, and S. Sinthupinyo, "Semi-supervised cluster-and-label with feature based re-clustering to reduce noise in Thai document images," *Original research article knowledge-based systems*, vol. 90, pp. 58-69, 2015.

[6] W. Innchum, "The classification automatic for Thai document by SVM and natural language processing," M.S. thesis, Dept. Engineering, Kasetsart University, Thailand, 2005.

[7] Yang, O. Perderson, "A comparative study on feature selection in text categorization," in *Proc. ICML-97, 14th international Conf. on machine learning*, 1997.

[8] P.R. Mahalingam, and S. Vivek, "Predicting financial savings decisions using sig moid function and information gain ratio," *in Proc. 6th international conference on advances in computing and communications*, pp. 19-25, 2016.

[9] J. Niwes, S. Parinya, and M. Payung, "Experimental studies reduction techniques features and algorithms for classification of Thai documents," *Science Journal King Mongkut's institute of technology the military Lat Krabang*, vol. 2, 2013.

[10] J.R. Quinlan, C4.5: Programs for machine learning. Morgan Kaufmann publishers, 1993.