PROGRESS IN APPLIED SCIENCE AND TECHNOLOGY
Faculty of Science and Technology
Rajamangala University of Technology Thanyaburi
Vol.10 No.02 (2020): 7–13
ISSN (Print): 2730-3012
ISSN (Online): 2730-3020
https://ph02.tci-thaijo.org/index.php/stj-rmutt

Received 10th September 2020,
Revised 14th October 2020,
Accepted 18th October 2020

DOI: 10.14456/past.2020.8

Sentiment Analysis of Thai Online Product Reviews using Genetic Algorithms with Support Vector Machine

Rawisuda Tesmuang[1*], Nivet Chirawichitchai[1]
[1] *Faculty of Information Technology, Sripatum University, Jatujak, Bangkok, 10900, Thailand*
*E-mail: Rawisuda35@gmail.com

Abstract

This research purposes sentiment analysis of Thai online product reviews for hotel room services, hotels, and resorts with a collection of 4,000 sample data sets. A Modeling with Genetic Algorithms with 4 machine learning methods is created. It consists of Support Vector Machine, Decision Tree, Naïve-Bayes, and K-Nearest Neighbor to compare the effectiveness of each method in analyzing sentiment analysis of the online products. The experiment found that the use of Genetic Algorithms with support vector machines provide better classification accuracy than using vector support machines with an accuracy of 88.64% and the proposed model can effectively reduce the dimensions of the data.

Keywords: Genetic Algorithms, Sentiment Analysis, Support Vector Machine

1. Introduction

Websites and social media play an increasingly important role in disseminating information nowadays, including access to review various services regarding online products, websites, hotel room service, hotels, and resorts. The Agoda Thailand website and Booking.com are very popular websites for hotel room service, hotels, and resorts. There are a lot of people who use them to book a hotel room and resort including commenting or review services. In addition, Social Media is a form that plays a very important role in broadcasting directly through Smartphones or Mobile devices, which supports a large number of users. It also helps the website owner getting access to reviewed messages from their customers. The researchers, therefore, use 4,000 data sets from product reviews collected from both Agoda Thailand and Booking.com by doing sentiment analysis of Thai Online Product Reviews to make it more convenient for future customers who use or are deciding to use the service. Future customers can read reviewed comments in order to make a decision that suits their preferences (2, 3). From the background above, the researchers have an idea of adding value and creating benefits for Thai Online Product Reviews by classifying the reviews based on Support Vector Machine, Decision Tree, Naïve-Bayes, and K-Nearest Neighbor techniques to test the model performance to suit the background of this research. It will also add value and benefits for customers who use or are deciding to use the service (1-3).

2. Theory and Related Literature

2.1 Text Mining

Text mining or Knowledge - Discovery is a process that deals with messages (commonly used with a lot of texts) to find patterns, approaches, and relationships hidden in that thread. This technique uses information retrieval, data mining, machine learning, statistics, and computational linguistics. The text mining technique is similar to data mining but the data mining technique is often used with data or structured databases. The text mining technique focuses on unstructured or semi-structured texts (4).

2.2 Genetic Algorithms (GAs)

Genetic algorithms or GAs is a mathematical model by Holland (5-6) a pioneer who discovered this model. This model is the best method for finding answers by using natural selection and species principles. The genetic algorithm is one of calculation that evolve in the process of finding answers. It has been classified as one in the group of Evolutionary Computing which is currently recognized for efficiency and is widely applied to solve the optimization problem. The genetic algorithm is a process of finding answers to the system. It acts as a tool in calculating. The Cycle of Genetic Algorithms basically consists of 3 major key

processes, as shown in figure 1 and explain process in the table 1.

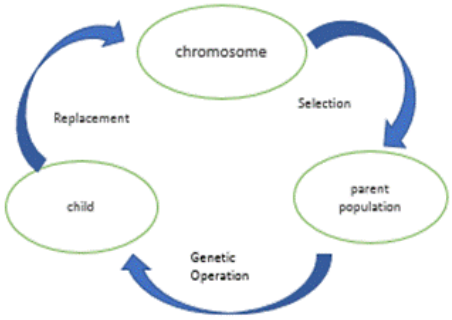

Figure 1. Cycle of Genetic Algorithms

Table 1 Cycle of Genetic Algorithms

Process No.	GAs process features
1) Selection	This is a step in selecting a good population in the system to be the origin of species and give birth to the next generation.
2) Genetic Operation	This process is a chromosome transforming process by means of species. It creates an offspring which is obtained from combining the origin of species that has combined from parents or by changing the genes of parents in order to get new offspring.
3) Replacement	This process is bringing the new offspring to replace the old generations. It Is a process in choosing which group of new generation and how many of them will be used to replace which group of the old generation.

The genetic algorithm simulates the evolution of life in natural systems, that is to say, the process inside a genetic algorithm makes the answer existed in the system evolve in itself. This process leads to adaptation and becomes a better and best answer. From figure 1, it can be seen that the components in the Genetic cycle, the algorithm consists of population, origin, species, new species, and details of each element.

2.2.1 Genetic Algorithm Procedures

General procedures of genetic algorithm and connecting to real-world systems to search for the desired answer. The answer that the system requires genetic algorithms to search is in chromosome forms in the population group (the desired answer must be the best chromosome in the group). Therefore, the system will be able to know whether the answers contained in the genetic algorithm at a given time are good or bad by evaluating the chromosomes. The system will connect to the genetic algorithm through the objective function to evaluate the chromosomes for each procedure shown in the pseudocode below.

```
Comment P is Population
 Initialize P
  While not terminate
  Evaluate P for fitness
  P' = Selection.Crossover.Mutation of P
  P = P'
  Terminate
```

1. Answers close to the target (less than £; referring to the minimum amount before a new cycle of the order amount is refilled)

2. The number of cycles exceeded the limit

From the pseudocode, you will see that the iteration will begin repeatedly from step 2 (Population Evaluation) until the desired answer is obtained. The answer will come from the best chromosome in the group of population. The values from objective functions can be used in order to assess whether the answer is needed. The GAs features of each procedure is described in table 2.

Table 2 GAs features of each procedure

Process No.	GAs features of each procedure
1) Population	Generate initial population which usually creates randomly
2) Population Evaluation	Evaluate the population or chromosome of the entire population using an objective function due to the system is unable to understand the value of the chromosome within the genetic algorithm. Therefore, the chromosome must be decoded before being calculated with the objective function.
3) Fitness	Calculate the suitability then return to the genetic algorithm.
4) Selection	Select by the suitability of certain chromosome groups to be used as the origin of the species. These species will then be used as a substitute for the relay breed for the next generation.
5) Genetic Operation	Operate the crossover and mutation process by bringing the origin of species to create offspring with species operations. The chromosome obtained at this stage is the offspring chromosome.
6) Replacement	Replace the original population's chromosome with the offspring obtained from item 5. Some of the population will be replaced by a specific strategy. The replacement process has used the appropriateness of the decision.

2.3 Naïve-Bayes

Naïve-Bayes is based on Bay's rules in probability theory and statistics to assess uncertainty into numbers of the probability of an event (A). If another incident has already occurred (B) can be written in a simplified form, as shown in equation 2.1 (7).

$$P(A|\mathrm{B}) = \frac{P(A \cap B)}{P(B)} \quad (2.1)$$

P(A|B), i.e. probability of event A will occur given that event B has already occurred

P(B|A), i.e. probability of event B will occur given that event A has already occurred.

P(A), i.e. probability of event A will occur

P(B), i.e. probability of event B will occur

To use probability principles to classify the categories is called the classification method. It helps predict and explain the results. It also analyzes the relationship between variables to use in creating the probability conditions for each relationship. This method is one of the effective methods to use for text classification. It is not complicated and suitable for large sample sets and the attributes are independent of each other. The probability of the data is defined as in equation 2.2

$$P(A_1, A_2, \dots, A_n|C_j) = \prod_{i=1}^{n} P(A_i|C_i) \quad (2.2)$$

Cj group, i.e. data that has n attribute X = {A_1, A_2, ..., A_n} or use the symbol as P = (A_1, A_2, ..., $A_n|C_j$), where Π represents all multiply result of P = (Ai|Cj) i = 1, 2, 3, ..., n and j = 1, 2, 3, ..., n, so we will get a simple Bayesian classification methods as shown in equation 2.3.

$$V_{NB} = argmax\ \mathrm{P(C_j)} \prod_{i=1}^{n} P(A_i|C_i) \quad (2.3)$$

2.4 Decision tree

The decision tree (8) is learning by doing data set classification. The data set is divided into various classes by using the attribute of the data to classify it. The decision tree obtained from learning to give us the knowledge of which data attribution determines the classification and how important each data quality is. This is useful for users to be able to analyze data and make more accurate decisions. The result of learning the decision tree consists of 1. Internal Node, 2. Branch or Link is the value of the attribute on the internal node that branched out. The internal nodes will branch into equal amounts of the attribute values of that internal node, 3. Leaf Node is a class of results in data classification. This process is used to create a decision tree by using a measuring value, called the Gain, to decide which attribute to use to divide the data for decision making. Determining the decision tree structure uses the order of gain value of the highest attribute. It will start finding the information value of each attribute that indicates the ability of the attribute to separate each class as equation 2.4

$$I(S_1, S_2, \dots, S_n) = -\sum_{i=1}^{n} \frac{Si}{s} log_2 \frac{Si}{s} \quad (2.4)$$

Where S is the data total which has S_i amount of $CLASS_i$

n is the class total. Then find the Entropy value of the sum of each attribute A that can classify information each class of each attribute value m number as in equation 2.5

$$E(A) = \sum_{j=1}^{m} \frac{S_{1j}+\cdots+S_{nj}}{S} I(S_{1j} + \cdots + S_{2j}) \quad (2.5)$$

Therefore, the gain value of the selected attribute can be found by classification of all the information of that attribute as in equation 2.6

$$Gain(A) = I(S_1, S_2, \dots, S_n) = -E(A) \quad (2.6)$$

2.5 K-Nearest Neighbor

This is a method of classifying classes. The principle of this method (9) is to classify data based on the nearest k members from the work sample set based on the smallest distance of the new member, input query instances with sample data, and training samples. It will calculate the nearest neighbor k data. After that, it will gather the nearest k members and choose the class that the most k members are into new members. The data classification using k-nearest neighbor consists of a multi-variable attribute Xi, which is used to divide the Yi group by specifying the positive integer value for k. This value indicates the number of the case to be searched in the prediction of the new cases. The KNN algorithms include 1-NN, 2-NN, 3-NN, …… k-NN, where the k value must be specified when creating the model (9-10)

The distance measure is finding the distance or length between the desired points by using various tools and methods. In this research, Euclidean Distance has been chosen as a method of finding an accuracy distance between 2 points. The point to be measured has conditions and many values from many dimensions or sizes depending on the model. It can be proved by Pythagorean's theory when the formula is used to find the Euclidean distance. The distance between points P = (p1, p2, ..., pn) and Q = (q1, q2 ..., qn) in Euclidean of many sizes can be specified as in equation 2.7.

$$\sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \tag{2.7}$$

2.6 Support Vector Machine

The theory is presented by Cortes and Vapnik (1995) to minimize errors from the prediction process. It is a technique used to solve data pattern recognition problems based on data classification by finding the decision plane and dividing the information into 2 parts. It tries to create the midline between the groups to have the most distance optimal separating hyperplane (9-11) to find the decision plane for data dividing. It tries to create a dividing line between the groups to have the most distance between the boundaries of both groups by using the mapping function to move data from the input space to the feature space and creating a similarity measure function called the kernel function. In the media about SVM we will variable used to make a decision as attribute and variable. They are used to determine a multi-dimensional plane called the feature. The most suitable selection is called a feature selection. The number of sets of features that are described in one instance (such as the row of predictive values) is called vector. Therefore, the purpose of the SVM model (7) is to get the most out of the multi-dimensional plane that separates groups of vectors. In this case, the feature space is suitable for data that has a given dimension of data the most where (xi, yi), ... , (xn, yn) is an example used for teaching, n is the amount of data, m sample is the number of input dimensions and y is the result with + 1 or -1 as in equation. For linear problems, high dimensions are divided into 2 groups by the decision plane which can be calculated as in equation 2.8.

$$(w * x) + b = 0 \tag{2.8}$$

Where w is the weight value and b is the bias value. The equation is used to classify the data as in equation 2.9.

$$\begin{aligned}&(w * x) + b > 0 \text{ if } y_i = +1 \text{ and}\\ &(w * x) + b < 0 \text{ if } y_i = -1\end{aligned} \tag{2.9}$$

In the sentiment analysis of Thai online product reviews that we collected using 4 genetic algorithms with support vector machine, Naïve-Bayes, decision tree, and K-Nearest Neighbor to compare the effectiveness of Thai sentiment analysis regarding online product reviews.

2.7 Related literature

The researchers applied the data mining method (11 - 14) which extracted the comments of electronic products and divided the comments into positive and negative reviews according to the type of products and classified the semantic categories of words that express the popular positive and negative 1, 2, and 3 words. Then took those positive and negative comments to teach with advice sentences from a recommender whether the type is positive or negative by using the simple Naïve Bayes method. After that, it showed the result in either positive or negative probabilities. The disadvantage of this research is, if the information used in teaching the weight of the sentences is given incorrectly, the result will also be wrong, as well as this research (13). The data mining has been conducted, based on opinions from documents related to research articles. We searched from electronic databases and websites. In this research, we used the support vector machine technique and compare it with a simple Bayes method and other techniques (14). This is different from other researches that have done sentiment analysis on hotel domains by summarizing the results of ideas categorized by characteristics: service, food, hotel condition, location, room, and etc., in order to increase the convenience for users who want to know only one information. Whereas, users can select the desired feature and choose a hotel to compare the features chosen which hotel is better. The information will be displayed in a bar graph comparing positive and negative comments from all comments obtained from the natural processing language. The natural processing language separates the words from the expression of words stored in the database. In addition, this research does not show detailed recommendations regarding the comment of each user which is necessary to display all comments. Therefore, our research has presented an analysis of recommendations from users from the hotel domain.

Our research has conflicts with the said research because we compared the data mining technique using the decision tree, the simple Naïve Bayes, support vector machine, and K-Nearest neighbor to summarize negative and positive comments. Then compare it with the point given by users who used that particular service. This is the conflict that we have with the above research. In this research, we are able to summarize the overview of the sentiment analysis of Thai online product reviews or services whether they are positive or negative. The information will be shown in the text by graphing all the comments to make it easy to analyze the Thai sentiment regarding online product reviews.

3. Research methods

The method of this research aims to develop a sentiment analysis of Thai online product reviews from Thai public resources. The conceptual framework for the development of sentiment analysis of Thai online product reviews as in figure 2.

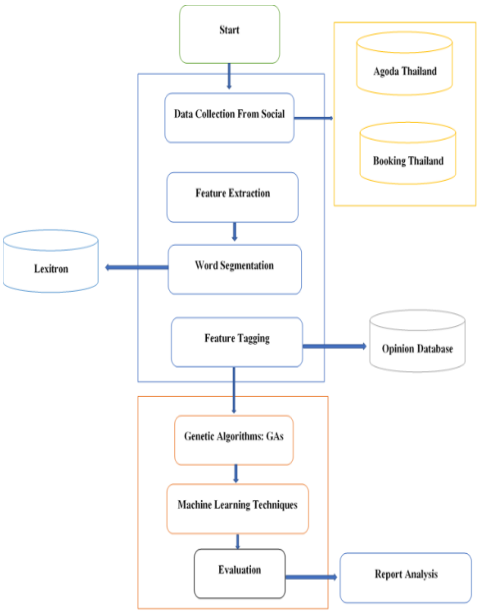

Figure 2 Conceptual framework of sentiment analysis of Thai online product reviews using genetic algorithm together with support vector machine

3.1 Data preparation

This ss a process of collecting sentiment analysis of Thai online product reviews that are used to create 4000 learning models from Agoda Thailand and Booking.com. The public news sources studied are as shown in Table 3.

Table 3 Volume of sentiment analysis of Thai online product reviews

Thai public news sources	Number of messages
positive	2000
negative	2000

3.2 Document Parser

The process of extracting data from data sources then filtering the data by eliminating duplication. The datum were conversed into the appropriate format for modeling by using text to break words through the Thai word processing program called Thai Lexeme Tokenizer: LexTo. Then tag the information to use for the training set and specify the categories of text by hand as in Table 4.

Table 4 Data tagging for the Training Set

Detail	Class
การ | บริการ | ดี | | พนักงาน | ดู | น้อย | ไป | หน่อย | ส่วน |โรงแรม | ดู | สะอาด | แต่ | ก็ | ยังมี | สิ่งอำนวยความสะดวก | ใน |ห้อง | ยัง | น้อย	Nagetive
บริการ|แย่มาก|จอง|เตียง|ใหญ่|เตียง|เล็ก|พนักงาน|เช็คอิน| ช้า|ไม่มี|เตียง|จองไว้|พนักงาน|พูดจา|แย่มาก|ลูกค้า|เสีย| แรง|จอง	Nagetive
เป็น | โรงแรม | ที่ | ดีเยี่ยม | มากๆ | คะ | สระ | ว่าย น้ำ | สะอาด สวยงาม | ห้องพัก | สะอาด | กว้างขวาง	Positive
บรรยากาศ | ดีมาก | วิว | ทะเล | สวย | มาก | สระ น้ำ | ดีมาก คน | ไม่ | เยอะ | ห้อง | หรู | | ไฟ | สว่าง	Positive

3.3 Feature extraction

The purpose of the feature extraction procedure is to extract the feature of the comments. The feature extracting should determine what is the representative of comment feature and which value will represent that particular comment feature. Then use the value representing that comment feature by creating the Tokenize words and filter the tokenize words by specifying words that are between 2 to 25 characters long and filling in some stop words which is the removal of insignificant words without changing the meaning of the news. At this stage, Rapidminer program is used to help extract the feature of sentiment analysis of Thai online product reviews as in figure 3.

Figure 3 Extraction sentiment analysis of Thai online product reviews by Rapidminer (9-10).

3.4 Stop-Word List Removal

This is to remove insignificant words without changing the meaning of the document. The insignificant words are words commonly used has no significant meaning to the document when they are being removed from the document, they do not change the meaning. For example, prepositions or words that connect words or groups of words together, conjunctions or words that connect words with other words, and pronouns or words that are used in place of nouns that have already mentioned in the sentence. Therefore, stop-words are considered insignificant to classify (9-10).

3.5 Stemming

This process is finding the root of a word or looking for words with similar meanings in order to combine them into one word. The stemming process should be done before indexing. It helps reducing the index size and increasing the efficiency of searching or classifying e.g. the stemming using Porter Stemming Algorithm (9-10).

3.6 Indexing

Due to computers are not able to directly classify documents in natural language, the documents should be converted to suit the computers' capability. The documents converting process is called indexing. This process creates a document representative to use in the learning process. The purpose of creating the indexing is to calculate values that will be used as document attributes. The indexing process commonly used to begin creating a representative vector document and matrix of the document groups created from all of the document vectors in the group (14-15). This research uses an experiment to assign weight values to the following index.

f_{ik} = frequency of the word *i* in document *k*

N = total number of documents of all groups

n_i = all document total where the word *i* occurring

Term Frequency–Inverse Document Frequency (TFIDF) This value is determined by the word frequency in the document multiplied by the log function of all documents and divided by the number of documents of that word occurring. This is a popular weight standard method as equation 3.1.

$$a_{ik} = f_{ik} * log(\frac{N}{n_i}) \quad (3.1)$$

3.7 Feature Selection

The feature selection is the reduction of data size by reducing the original data size and losing important characteristics the least. There are different techniques of feature selection to achieve different characteristics as well. The feature selection has many names but in the statistical field, it's often called variable selection because each attribute is considered as random variables. Sometimes, the variable is called subset selection because it selects a certain number out of all Is the subset selection. In this research, we used the genetic algorithm to reduce data characteristics (12-13).

3.8 Model Builder

This is to create modeling for learning using 4 classification algorithms: Naïve Bayes as in figure 5, SVM as in figure 7, K-Nearest Neighbor as in figure 4, and the Decision Tree as in figure 6 to compare the classification performance.

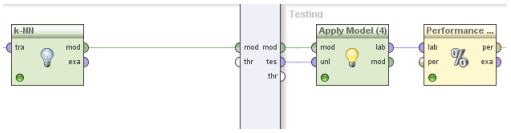

Figure 4 Modeling using the K-Nearest Neighbor algorithm

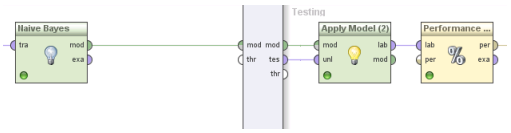

Figure 5 Modeling using the Naïve Bayes algorithm

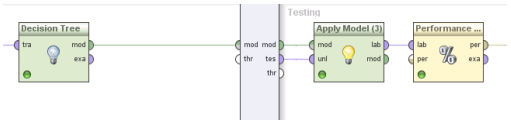

Figure 6 Modeling using the Decision Tree algorithm

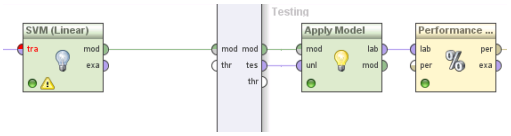

Figure 7 Modeling using SVM algorithm

3.9 Classifier

Classification of sentiment analysis of Thai online product reviews is divided into 5 categories: 1) cleanliness, 2) facilities, 3) comfort and room quality, 4) services, and 5) worth spending money.

3.10 Testing and Evaluation

Modeling of data classification test sentiment analysis of Thai online product reviews from Thai public resources. We considered the accuracy by using the ability assessment of the model to measure the effectiveness of data classification according to the concept of information retrieval which is the measurement of the accuracy as equation 3.2 (14-15).

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (3.2)$$

4. Model Performance

From the results of the sentiment analysis of Thai online product reviews using the genetic algorithm to reduce the size of the data dimension. The result from the data dimension being reduced was being sent into the machine with effective learning solutions and conducted a comparative test of accuracy performance. When substituting the index values using the TFIDF method by 4 learning algorithms, it was found that the use of genetic algorithm together with the support vector machine provides the most classification efficiency of 88.64%, followed by using the support vector machine only gave the classification efficiency of 87.17%, the K-Nearest neighbor gave the classification efficiency of 85.51%, the genetic algorithm together with the K-Nearest neighbor gave the classification efficiency of 79.97%, the genetic algorithm together with the decision tree gave the classification efficiency of 67.18%, and the genetic algorithm together with Naïve-Bayes gave the classification efficiency of 77.49% respectively as in the figure 8.

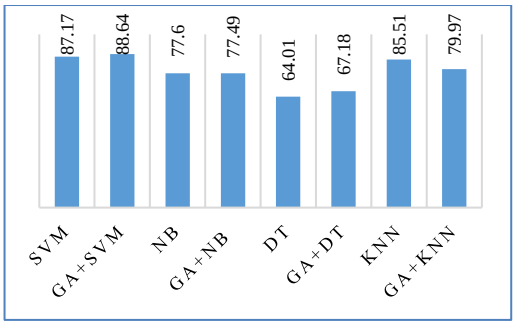

Figure 8 Test results comparison of each algorithm

5. Summary

In this research, the researchers present a sentiment analysis of Thai online product reviews using the genetic algorithm with Support Vector Machine from Agoda Thailand and Booking.com and create a model by using 4 algorithms: Support Vector Machine, Naïve-Bayes, Decision Tree, K-Nearest Neighbor in order to compare the accuracy values. We found that to reduce the characteristics with the genetic algorithm and a learning machine using the genetic algorithm together with the Support Vector Machine gave the best classification efficiency. It gave the highest accuracy of 88.64%, which will provide a good classification of feedback written from the feelings and emotions of users. This type of feedback can reduce the data dimension using the genetic algorithm. From the proposed process, we found that the data dimension reduction does not affect the efficiency of the data classification in any way and it can also be applied to other services.

Declaration of conflicting interests

The authors declared that they have no conflicts of interest in the research, authorship, and this article's publication.

References

1. Katrekar A, AVP BD. An introduction to sentiment analysis. GlobalLogic Inc. 2005.
2. Pang B, Lee L. Opinion mining and sentiment analysis Foundations and Trends in Information Retrieval Vol. 2.
3. Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. arXiv preprint cs/0205070. 2002 May 28.
4. Janpla S, Wanapiron P. System Framework for an Intelligent Question Bank and Examination System. Int. J. Mach. Learn. Comput. 2018 Oct;8(5).
5. Mitchell M. An introduction to genetic algorithms. MIT press; 1998.
6. Sıvanandam SN, Deepa SN. Introduction to Genetic Algorithm. Springer Science & Business Media Publisher; 2008.
7. Khan K, Baharudin BB, Khan A. Mining opinion from text documents: A survey. In2009 3rd IEEE International Conference on Digital Ecosystems and Technologies 2009 . Jun 1 pp. 217-222.
8. Salzberg SL. C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993.
9. Chirawichitchai N. Automatic Thai Document Classification Model. J. Ind. Technol. 2013 Jan;9(1).
10. Chirawichitchai N, Sanguansat P, Meesad P. Automatic Thai Document Categorization with Support Vector Machines. InProceedings of the 6th National Conference on Computing and Information Technology 2010.
11. Joachims T. Text categorization with support vector machines: Learning with many relevant features. InEuropean conference on machine learning 1998 Apr 21. pp. 137-142.
12. Kongthon A, Angkawattanawit N, Sangkeettrakarn C, Palingoon P, Haruechaiyasak C. Using an opinion mining approach to exploit web content in order to improve customer relationship management. InPICMET 2010 Technology Management for Global Economic Growth 2010 Jul 18. pp.1-6.
13. Joachims T. Text categorization with support vector machines: Learning with many relevant features. InEuropean conference on machine learning 1998 Apr 21. pp.137-142.
14. Chirawichitchai N. Developing term weighting scheme based on term occurrence ratio for sentiment analysis. InInformation Science and Applications 2015 pp. 737-744.
15. Chirawichitchai N. Sentiment classification by a hybrid method of greedy search and multinomial naïve bayes algorithm. In2013 Eleventh international conference on ICT and knowledge engineering 2013 Nov 20. pp.1-4.