ORIGINAL ARTICLE SPECIAL ISSUE: Advanced Analyses of Technology Innovations, Patents, and Intellectual Property
Guest edited by Amy J. C. Trappey

Application of the Honeybee Mating Optimization Algorithm to Patent Document Classification in Combination with the Support Vector Machine

Chui-Yu Chiu* and Pei-Ting Huang
Department of Industrial Engineering and Management, National Taipei University of Technology, Taiwan
(**Received** 3 December 2012; **Accepted** 8 February 2013; **Published on line** 1 September 2013)
*Corresponding author: cychiu@ntut.edu.tw
DOI: 10.5875/ausmt.v3i3.183

Abstract: Patent rights have the property of exclusiveness. Inventors can protect their rights in the legal range and have monopoly for their open inventions. People are not allowed to use an invention before the inventors permit them to use it. Companies try to avoid the research and development investment in inventions that have been protected by patent. Patent retrieval and categorization technologies are used to uncover patent information to reduce the cost of torts. In this research, we propose a novel method which integrates the Honey-Bee Mating Optimization algorithm with Support Vector Machines for patent categorization. First, the CKIP method is utilized to extract phrases of the patent summary and title. Then we calculate the probability that a specific key phrase contains a certain concept based on Term Frequency - Inverse Document Frequency (TF-IDF) methods. By combining frequencies and the probabilities of key phases generated by using the Honey-Bee Mating Optimization algorithm, our proposed method is expected to obtain better representative input values for the SVM model. Finally, this research uses patents from Chemical Mechanical Polishing (CMP) as case examples to illustrate and demonstrate the superior results produced by the proposed methodology.

Keywords: Honey-Bee Mating Optimization; Support Vectors Machine; TF-IDF; Document Categorization

Introduction

With the advance of information technology, foreign scholars and entrepreneurs have gradually been paying increasing attention to text mining. For example, M. Fattori [1] proposed text mining technology for patent analysts that could overcome patent category limits. Text mining techniques and patent classifications should not be considered alternative tools for patent mapping. Rather, they should be used in synergy. According to the World Intellectual Property Organization (WIPO), a patent is a unique knowledge document disclosing the core technology of an item. The ability of a patent to transmit core technology is much higher than that of a journal or a research report; therefore, the patent itself can be the source of technology prediction, and the annual quantity of patents being issued can be regarded as the share of the “knowledge market".

This study used a chemical mechanical polishing (CMP) patent knowledge document as the carrier of the system, in order to test the performance of the system. CMP is a planarization technology extensively used in the semiconductor process. The development and application of patents in this area are very important, and the proper classification of documents in this domain will be helpful for improving the accuracy and speed with which documents can be searched for and located as well as for avoiding patent traps. This study used the International Patent Classification (IPC) process as the patent classification criteria for performance evaluation.

Patent classification is combined with text mining for the retrieval of patent documentation. As different researchers carry out data retrieval for different subjects, there will be variations due to the researchers' experience, thinking logic and accumulated knowledge. For example, the IPC has 69,000 main items and sub-items. How to choose appropriate patent keywords and patent categories to attain a high accuracy and recall

rate for patent documentation in the text mining process is an important subject of research for development personnel and inventors.

This study aimed to construct an improved classification method by adding the improved characteristics of the honey-bee mating optimization algorithm proposed by Haddad with a support vector machine, in order to correct the parameters for patent document classification and to generate a high performance classifier capable of increasing the accuracy rate of document classification.

Literature Review

Support Vector Machine

The support vector machine (SVM) is a machine-learning algorithm that was proposed by Vapnik et al. [2] based on statistical learning theory. Salton [3] published the concept of SVM at the ACM international seminar in January 1964, and he proposed a vector space architecture in 1975. Later Boser et al. [4] proposed a direction to establish a nonlinear classifier, i.e. using the kernel function skill in the maximum margin hyperplane . Finally, Vapnik et al. [5] proposed support vector regression machines (SVR), so that the SVM could be applied to statistics.

The key concept of this study was to determine the maximum margin of separation on both sides of the hyperplane and to divide the data into two classes. The methods for the SVM processing of linearly separable cases, linearly inseparable cases and non-linear cases are introduced below.

Linearly separable cases

First, set S, which is formed of l training sample, is given, in which the training samples are (x1,y1),…,(xl,yl), and where x represents the training sample ($x_i \in R^N$,i=1,…l), y represents the class of the training sample, and its value is -1 or +1. SVM is used to correctly determine a hyperplane that can be separated into two classes of samples and to maximize the distance between the two classes. This hyperplane can be expressed as the following functional equation:

$$f_H(x) = w \cdot x + b. \quad (1)$$

Chui-Yu Chiu is an associate professor in the Department of Industrial Engineering and Management at National Taipei University of Technology, Taiwan. He received his Ph.D. degree in the Department of Industrial and System Engineering from Auburn University. His research interests include data mining with applications and intelligent management systems.

Pei-Ting Huang received her master of science degree in the Department of Industrial and System Engineering from National Taipei University of Technology. Her research interests include data mining and patent document categorization.

where w and b are derived from the training samples. The data classification result is verified by this hyperplane. A discriminant functional equation is constructed according to the hyperplane:

$$f_D(x) = sign(w \cdot x + b) = \begin{cases} +1, & y_i = +1 \\ -1, & y_i = -1 \end{cases}. \quad (2)$$

The discriminant is normalized to make all the samples of the two classes meet $|f_H(x)| \geq 1$. The maximum spacing between the two classes is equal to:

$$D(w,b) = \min_{\{x,y=1\}} \frac{w \cdot x + b}{\|w\|} - \max_{\{x,y=-1\}} \frac{w \cdot x + b}{w} = \frac{2}{\|w\|}. \quad (3)$$

In other words, the maximum interval is the minimized $\|w\|$ (or $\|w\|^2$). Therefore, the following equation should be met in order to obtain the correct classification of all the samples:

$$y_i(w \cdot x + b) \geq i = 1,....,l . \quad (4)$$

The plane meeting the aforesaid two conditions is called the optimal hyperplane. The samples in (4), in which the equal sign is tenable, are called support vectors.

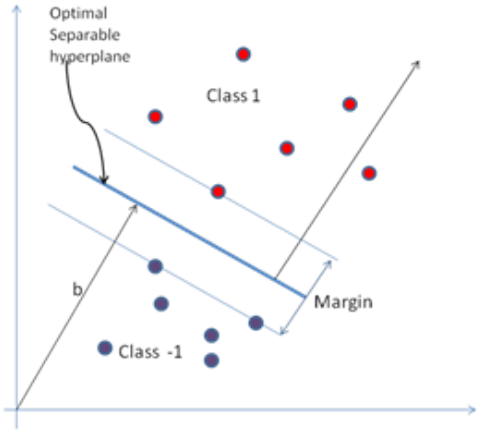

Figure 1. Maximum margin interval of SVM [4].

As mentioned above, the optimal hyperplane problem can be converted into a linear optimization problem, i.e.:

$$\begin{aligned} &Minimize \ \frac{1}{2}\|w\|^2 \\ &Subject \ y_i(w \cdot x + b) \geq 1, i = 1,...l \end{aligned}. \quad (5)$$

This is a quadratic programming problem that can be solved by the Karush-Kuhn-Tucker condition. First, the following Lagrangian function is defined:

$$L(w,b,a) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} a_i(y_i(w \cdot x + b) - 1), \quad (6)$$

where a_i represents any Lagrangian multipliers that are larger than zero.

The partial differential of $L(w,b,a)$ are shown as Equation (7) and (8).

$$\frac{\partial L(w,b,a)}{\partial b} = \sum_{i=1}^{l} a_i y_i = 0, \tag{7}$$

$$\frac{\partial L(w,b,a)}{\partial a} = w - \sum_{i=1}^{l} a_i y_i x_i = 0. \tag{8}$$

This is converted into the dual problem in Equation (9).

$$Maximize \sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} a_i a_j y_i y_j x_i \cdot x_j, \tag{9}$$

$$Subject \begin{cases} \sum_{i=1}^{l} a_i y_i = 0 \\ a_i \geq 0,\ i = 1,\ \ldots,1 \end{cases} \tag{10}$$

Only the coefficient a_i of the support vectors is not 0, consequently only the support vectors will influence the final division result.

$$w = \sum_{support vectors} a_i y_i x_i. \tag{11}$$

In other words, the weight coefficient vector of the optimal hyperplane is the linear combination of the vector samples. If a_i^* is the optimal solution, the functional equation of the optimal hyperplane will be:

$$f_D(x) = sign(w^* \cdot x + b^*) = sign\left[\sum_{i=1}^{l} a_i^* y_i (x_i \cdot x) + b^*\right], \tag{12}$$

where the sign () is the symbolic function, b* is the threshold of classification, and w can be worked out by Equation (12) from any support vector or from the intermediate value of any pair of support vectors from two classes. For a given unknown sample x, as long as $sign(w \cdot x + b)$ is calculated, the class of x can be determined.

Linearly inseparable cases

For non-linear sample data, either the error classification points should be reduced or the optimum hyperplane should be obtained, in order to maximize the margin interval between the two classes. Herein, the nonnegative slack variable ξ is imported to alleviate misclassification. Therefore, Equation (4) is changed to:

$$y_i (w \cdot x + b) - 1 + \xi_i \geq 0, i = 1,\ldots,l. \tag{13}$$

Two rules are imported based on the cost concept for processing inseparable samples:

- Maximum boundary: As in the case for separable samples, the optimal hyperplane is expected to separate the maximum spacing distance between two classes.
- Least error: Add $\xi 0 \geq$ as a penalty provision in the inseparable samples. As shown in Figure 2, not all of the points can meet Equation (4).

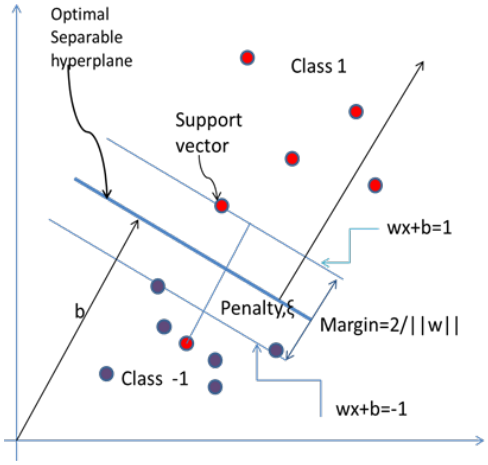

Figure 2. Penalty provisions for non-linear separation in SVM [4].

Therefore, polynomial $\sum_{i=1}^{l} \xi_i$ is regarded as the misclassification rate. This polynomial is imported into the inseparable condition, and Equation (5) will be changed to:

$$Minimize \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} \xi_i, \tag{14}$$

$$Subject \begin{cases} y_i (w \cdot x + b) - 1 + \xi_i \geq 0, \\ \xi_i \geq 0, \end{cases} i = 1,\ldots,l, \tag{15}$$

where *C* reflects the trade-off between the proportions of the complex and inseparable samples, which is also known as the penalty cost. The larger the *C* value , the more severe the conversion of the penalty into an error will be. The method of solving the optimization problem is similar to that for the separable condition, in that the Lagrangian multipliers are imported to convert it into a dual problem:

$$Maximize \sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} a_i a_j y_i y_j x_i \cdot x_j, \tag{16}$$

$$Subject \begin{cases} \sum_{i=1}^{l} a_i y_i = 0, \\ 0 \leq a \leq c_i, \end{cases} i = 1,\ldots,l. \tag{17}$$

Due to the Karush-Kuhn-Tucker condition,

$$a_i^*\left(y_i\left(w^*\cdot x_i+b_i^*\right)-1+\xi_i^*\right)=0, i=1,...,l, \quad (18)$$

$$(C-a_i^*)\xi_i^*=0, i=1,...l. \quad (19)$$

When $a_i^*\geq 0$, the sample x_i will be called the support vector. There are two patterns of inseparable support vectors, as follows:

(1) $0<a_i^*<C$: Same as for the separable sample, when slack variable $\xi_i=0$, $y_i\left(w^*\cdot x_i+b_i^*\right)=1$.

(2) $a_i^*=C$: When slack variable $\xi_i\neq 0$, the sample X_i that cannot meet Equation (4) will be regarded as an error.

Non-linear case

In addition to the linear sample described earlier, how to use SVM to solve classification problems when non-linear samples occur will be discussed below. In Figure 3, the left figure shows a sample where the linear hyperplane cannot be found for separation. The non-linear function Φ is used to linearly convert the originally non-linear sample space into a linear one, so as to find out a linear optimization hyperplane. In other words, the inner product of support vector s and sample vector x can be replaced by the kernel function, in order to map the space to a new optimal hyperplane [6].

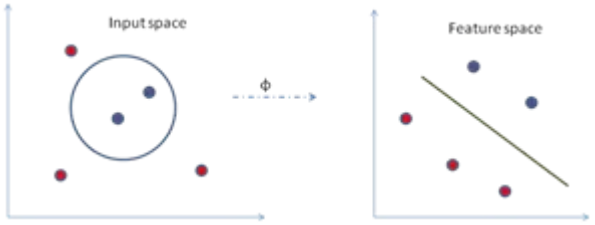

Figure 3. Converting the input space into the feature space.

The kernel function provides the efficiency of solving the non-linear separation problem. The kernel function (20) conforms to Mercer's theorem to meet an inner product that needs to be converted into high dimensional space [7].

$$K(x_i\cdot x)=\Phi(x_i)\cdot\Phi(x). \quad (20)$$

The kernel function is free from the complexity of calculating $\Phi(x_i)\cdot\Phi(x)$; therefore, the original Equation (9) is reduced to:

$$Maximize\sum_{i=1}^{l}a_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}a_ia_jy_iy_jK(x_i\cdot x), \quad (21)$$

$$\text{Subject}\begin{cases}\sum_{i=1}^{l}a_iy_i=0,\\ a_i\geq 0,\end{cases} i=1,...,l. \quad (22)$$

The optimal hyperplane can also be a low dimensional feature space, and the functional equation of the optimal hyperplane for this is:

$$f_D(x)=sign\left(w^*\cdot x+b^*\right)=sign\left[\sum_{i=1}^{l}a_i^*y_iK(x_i\cdot x)+b^*\right]. \quad (23)$$

The optimal hyperplane depends on the adopted kernel function. The kernel functions that conform to Mercer's theorem include the Gaussian kernel and the polynomial kernel [8-10], described below:

(1) Polynomial kernels of degree d

$$K(x,y)=\left[(x,y)+1\right]^d. \quad (24)$$

(2) Radial basic functions (RBF) with Gaussian kernels of width C > 0

$$K(x,y)=\exp\left(-\gamma\|x-y\|^2\right). \quad (25)$$

(3) Neural networks with tanh activation function

$$K(x,y)=\tanh\left[K(x,y)+\mu\right]. \quad (26)$$

Honey-Bee Mating Optimal Algorithm

HBMO Development Process

In recent years, the optimal algorithm has been widely applied based on evolutionism proposed by Darwin. Several algorithms have been found in natural phenomena which focus on the optimal solution for heredity operations including selection, replication and mutation. The optimal algorithm has extended to various areas such as data mining, biotechnology, finance and economics, and even the design of circuitry.

Marriage in Honey Bees Optimization (MBO) was proposed by Jason Teo and Hussein A. Abbass [11, 12]. Hussein brought up a swarming relation in society model assuming a polygynous colony. Each bee performs sequences of actions which accord to genetic, environmental, and social regulation. The result of each action influences the subsequent actions of both a single bee and many drones. The marriage process represents one type of action to construct the optimization algorithm.

Procedure of HBMO

The Honey-Bee Mating Optimal Algorithm (HBMO) is a hybrid evolutionary algorithm including Simulated Annealing (SA), and the Genetic Algorithm (GA). HBMO was developed by Bozorg Haddad [13] and Afsha [14].

The HBMO algorithm consists of the following five stages [11]:

(1) The algorithm starts with the mating flight where a queen (best solution) selects drones probabilistically to form the spermatheca (list of drones). A drone is then selected from the list randomly for the creation of broods.
(2) Creation of new broods (trial solutions) by crossover of the drone's genotypes with the queens.
(3) Use of workers to conduct local search on broods (trial solutions).
(4) Adaptation of worker's fitness, based on the amount of improvement achieved on broods.
(5) Replacement of weaker queens by fitter broods.

The procedure of the HBMO algorithm consists of the following duplicate characters:

a. Reproduction

In this stage, the new set of broods (trial solution) is generated from the crossover process by mating drones' genotypes with the queen's (best solution) randomly. The drone would be updated from the best list of drones or the newly generated set of solutions using simulated annealing to select the set of solutions from the search space.

b. Improvement

Improvement involves a newly generated set of solutions employing different heuristic functions and mutation operators according to their fitness values.

The HBMO algorithm integrates GA to generate the new broods, and SA to simulate the mating and selecting of drones.

In order to visualize the procedure of HBMO clearly, Figure 4 shows the HBMO flow chart and Figure 5 is the pseudo-code of the HBMO algorithm.

The predefined parameter is the number of workers (W) representing the number of heuristics encoded in the program. The queen's spermatheca size represents the maximum number of matings per queen. When all queens complete their mating flights, workers start to feed and mutate. After all broods have been generated, they are sorted according to their fitness. The best brood replaces the worst queen until there is no brood that is better than any of the queens. Remaining broods are then killed and a new mating flight begins until all assigned mating flights are completed or convergence criteria are met.

Application of honeybee mating optimization

The Honeybee Mating Optimization algorithm is an improved algorithm that combines the genetic algorithm with the simulated annealing approach [15]. The crossover and mutation concepts of the genetic algorithm are also introduced, so that the generation of solutions can jump off the local optimum. The simulated annealing approach uses a disturbance mechanism, and the generation of the solution also gets rid of the local optimum. The simulated annealing approach is characterized by a steep slope method that can obtain a solution faster.

The Honeybee Mating Optimization algorithm has been gaining popularity, and it has been used in a number of studies. For example, Afshar et al. [14] used HBMO to optimize the general situation of reservoir operation, solved non-linear and continuous restrictive problems, and validated the efficiency of the algorithm. Marinaki et al. [16] used HBMO in the first stage of a two-stage classification, and used the nearest neighbor to solve a financial classification problem. Niknam [17] used HBMO to evaluate a distribution system.

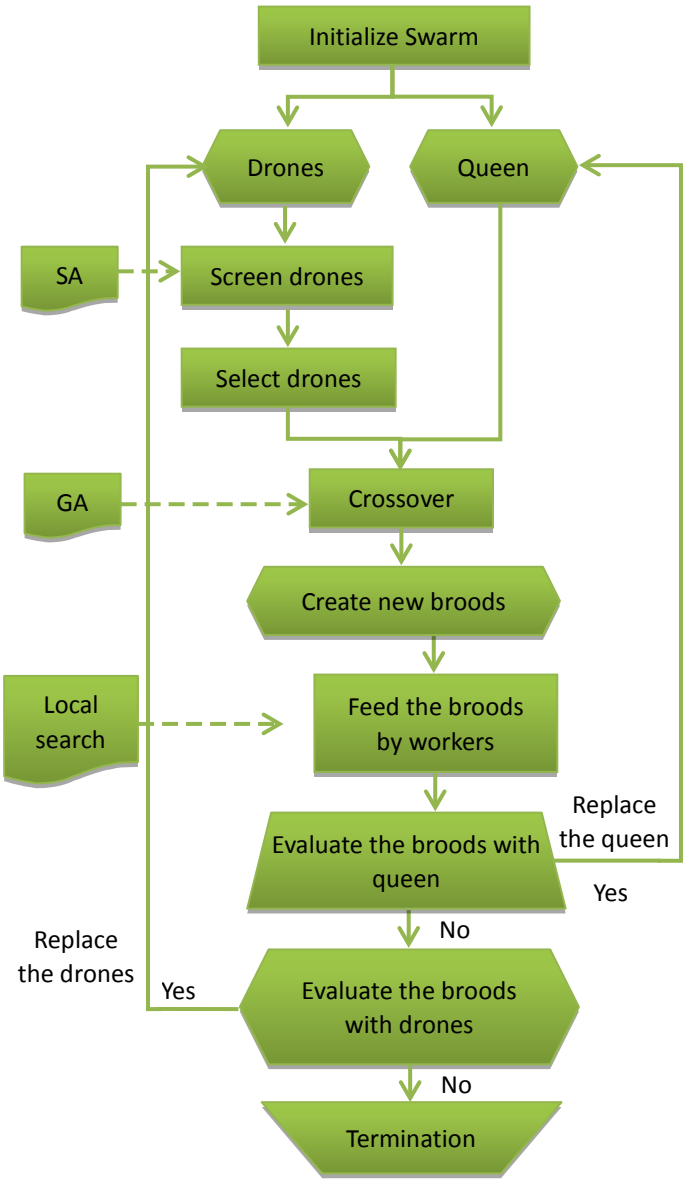

Figure 4. HBMO procedure [18].

Set H Number of bees in Hive
Set D Number of Drones in Hive
Set S Capacity of Spermatheca
Set T_{max} Speed of Queen at the start of a Mating Flight
Set T_{min} Speed of Queen at the end of a Mating Flight
Set t Speed reduction rate
Set MF Number of Mating Flights
Set W Number of Workers
Repeat for all H
Generate H_i
Calculate objective function ($f(H_i)$)
Queen (Q) = Best (H_i)
EndBegin
Repeat for all H_i
Generate r = (0,1) randomly,
If $\exp(-ABS(f(Q)-f(H_i))/T) > r$
Then (add Hi to spermatheca (S_i) And $T = t*T$) **Else** $T=T$
Until ($H=0$ or $i=S$ or $T=T_{min}$) **End**
Repeat for all W_j
Repeat for all Broods supposed to be produced (B_j) according to W_j Goodness probability ($g(Wj)$)
If Wj is a crossover function Then generate new Brood (Hi) by crossing over Si and Q
If W_j is a mutation function **Then** generate new Brood (H_i) by mutating S_i or Q
End
End
Calculate $f(H_i)$
Select Best (H_i)
If $f(\text{Best}(H_i)) > f(Q)$ **Then** Q = Best (H_i) **Else** $Q=Q$
Calculate Workers' goodness probability ($g(W_j)$)
Until $MF_k=MF$
Finish

Figure 5. the pseudo-code of the HBMO algorithm [13].

Text Mining

Knowledge discovery (KD) is a process for extracting implicit, useful, undiscovered and potentially valuable rules, information or knowledge [19]. The procedure of knowledge discovery consists of gathering data, purging data, converting data, applying mining technology, and presenting and interpreting the results. This procedure is similar to the patent analysis procedure, meaning that patent analysis is a form of knowledge discovery.

According to the differences in data features, knowledge discovery can be divided into data mining (DM) and text mining (TM). The former is used for structured data, meaning the data have common fields that can be recorded in a database, while the latter is used for unstructured data, meaning the data have no common structures; they may be in different lengths or consist of freely-recorded text information.

Mining procedures consist of association, classification, clustering, summarization, generation, prediction and sequence analysis. TM methods are related to the term frequency and the number of articles. TM is mainly applied to voluminous document libraries for data searches, knowledge extraction, decision aids, case tracking, crime analysis and trend predictions. Much attention has been paid to text mining technology in recent years. For example, R.J. Mooney and U.Y. Nahm [20] indicated that TM is a process often used for extracting modes from unstructured text.

Technology changes quickly and patent analysis has developed rapidly. Differing from the manual patent analysis of the past, information technology is presently being used in many countries to create patent maps or to analyze patent content. The Information Technology Center of the Taiwan National Science Council uses PatentGuider, which was developed by LearningTech Taiwan, and the Korea Scientific Information Center uses its PIAS series map for statistics and the mapping of structured field data in the patent context. HP uses Goldfire Innovator TM to innovate, research and develop products. Patent analysis software is applied to patent portfolios and patent attack scenarios. The functions of patent analysis include information retrieval, summarization, subject conclusion and classification, and the key point of such analysis is the recognition of patent keywords. However, keyword extraction still faces numerous problems and challenges.

Lo [21] established an automatic classification mechanism based on TM and SVM, which can filter customer opinions to improve the service efficiency of customer service units and customer satisfaction. Chang, Lin, & Wang [22] used content analysis and decision tree analysis for the TM analysis of customer related data, in order to classify customers and provide corresponding services for a more efficient market customer relation management strategy. Pons-Porrata, Berlanga-Llavori, & Ruiz-Shulcloper [23] proposed a new ascending hierarchical clustering algorithm that can cluster numerous news events according to their topics and can effectively shorten the user's topic search time. Delen & Crossland [24] used text mining technology to analyze the abstracts of theses in target journals, so as to find out the most frequent keywords by time interval, and to work out the differences among the emphasized topics of the same attribute in time intervals, thus allowing the users to find their information target rapidly. Hung, Chi & Chen [25] used self-organizing map technology to analyze user search behavior and store frequently used keywords in the interest library of the user's computer, as well as to dynamically update keywords data.

Text feature extraction methods

In high dimensional and sparse data sets, the efficiency of the classification and computing time are important, so in the machine learning document classification domain previously discussed, many articles have emphasized the main feature values only, instead of all the feature values. Thus, how to choose an appropriate feature extraction tool is an important goal.

At present, the generally accepted feature extraction methods are as follows:

(1) Term frequency-inverse document frequency (TF-IDF)

TF-IDF is a common weighted statistical method for information retrieval and document mining, which is used to evaluate the importance of a word to a file in a file set or a semantic library. The importance of the word is proportional to its occurrence frequency in a document, and it is inversely proportional to its occurrence frequency in the word library.

In a document, the term frequency ($tf_{i,j}$) often refers to the frequency occurrence of a word in a document, and it is normalized to avoid long documents. (The TF of the same word in a long document may be higher than that in a short document, no matter if the word is important or not). The importance of word ti in a document can be expressed as:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}, \quad (27)$$

where $n_{i,j}$ is the occurrence frequency of the word in document dj.

Inverse document frequency ($idf_{i,j}$) refers to the universal importance of words. The IDF of a word can be a logarithm of the value obtained by dividing the total number of documents by the number of documents containing the word:

$$idf_{i,j} = \log\left(\frac{N}{n_t}\right), \quad (28)$$

where N is the total number of documents in the semantic library, and n_t is the number of documents containing word ti.

The IDF can be used to distinguish the feature representativeness between documents. TF is combined with IDF as follows:

$$tfidf_{i,j} = tf_{i,j} \times idf_{i,j}. \quad (29)$$

(2) Chi-square statistics

The chi-square statistics test the independent deficiency degree of word j for class c. The larger the value is, the more important the word will be to the class. The equation is:

$$X^2(j,c) = \frac{N \times (AD - BC)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)}, \quad (30)$$

where N is the total number of documents in class C.

This can be described as shown in Table1. A is the number of articles containing word j in class C. B is the number of articles not containing word j in class C. C is the number of articles containing word j, however, not in class C. D is the number of articles not containing word j, and, not in class C.

Table 1. Distribution list of articles containing word j in class C.

		Word *j*	
		Containing	Not containing
Class C	**Yes**	A	B
	No	C	D

(3) Correlation coefficient

According to the findings of H. T. Ng [26], when positively correlated and negatively correlated words occur at the same time, the chi-square statistics will show them as positive values; therefore, the articles of occurrence and nonoccurrence and the words in class C will be selected as the feature values of class C. As it is not helpful for document classification, scholars have advocated using the correlation coefficient, i.e., using the unilateral chi-square statistics to choose words.

$$C_0(j,c) = \frac{N \times (AD - BC)}{(A+C) \times (B+D) \times (A+B) \times (C+D)}. \quad (31)$$

Research Method

SVM is a method for processing sample classification problems, and it is has been extensively used in various domains, such as patent analysis, catechism mining, e-mail virus recognition, handwriting recognition, image classification, and bio information. The clustering methods used in TM classification problems are the heuristic decentralized algorithm, K-cluster clustering, self-organizing maps and the EM algorithm. SVM is characterized by processing a high dimensional sparse matrix and having fewer control parameters. In the past SVM used a grid search algorithm to predict the two major parameters, the slack variable and the penalty cost in the model during calculations; however, this was very likely to cause a local optimum. The scholar Haddad [13] proved that the HBMO algorithm is better than both the genetic algorithm and the simulated annealing approach. The algorithm uses a queen and drone mating optimization-based screening mechanism that can provide better quality in local searches. Therefore, this study proposed using HBMO to solve the parameter adjustment, so as to enhance the classification effect of the classifier used. Therefore, the key point was the adjustment of the SVM parameters.

This study determined the optimum parameters by using HBMO based on SVM, in the hope of providing a more effective classification. This study used a chemical mechanical polishing (CMP) patent document as the experimental sample and used a hybrid algorithm combining HBMO with SVM to solve the patent classification problem.

The Case analysis flow chart is shown in Figure 6. The procedure of this research method is described below.

Stage 1: Data preprocessing

(1) Patent document collection

The data source of this project was mainly from Taiwan's patent data group. Between January 1998 and December 2009, there were approximately 860,000 items of patent data. As the classes of the various patent data needed to be defined prior to the training process of automatic classification, this paper used CMP as the main target data set.

(2) Auto tag

Regarding the auto tag function, the word set is generally derived from semantic analysis. An article is divided into words and the parts of speech of various words are marked. As semantic analysis is complicated, this experiment used the auto tag tool CKIP (Chinese Knowledge and Information Processing) developed by the Chinese word library team of Academia Sinica to tag words in the documents, as well as to automatically tag the patent title and summarization. In order to reduce the complexity of text feature processing, only two major word classes, the noun set and the verb set were extracted from the tagged words.

Stage 2: Feature extraction

The procedures for feature extraction and feature value calculation were as follows:

(1) The auto tag software tags the patent title and patent summarization in the data set collected by the aforesaid data gathering method, and then lists the results.
(2) The word feature values are calculated by using chi-square statistics, and the first n words of the class are extracted in turn.
(3) Finally, a high dimensional sparse matrix table of each patent and the TFIDF value of the extracted critical feature is calculated as the input vector of the main classifier.

Stage 3: HBMOSVM

This study used RBF as the kernel function for SVM, because the RBF kernel function can deal with high dimensional data problems. The adopted parameters were C (cost) and γ (slack variables), and the parameters and feature values derived from the HBMO algorithm were used. The detailed procedure for generating the optimum parameters and feature values were as follows:

(1) Define the input parameters for the model, the quantity of queens, the size of the storage capsule, and the quantity of house bees required.
(2) Randomly generate the initial drone.
(3) Perform scaling. Scaling can avoid the attributes of larger values influencing the attributes of smaller values, and can reduce computation complexity. The scaling of feature values can increase the accuracy rate of SVM. Generally speaking, the numerical range can be defined as [-1,+1] or [0,1] after scaling.
(4) Sequence according to the fitness values. Keep the best solution as the queen, and keep the rest as drones.
(5) Use the simulated annealing approach to screen the drone swarm for mating, and then replace the present optimal solution and trial solution.
(6) Generate new feasible sets according to the crossover and mutation mechanisms of the genetic algorithm, and then mark them as the queen (the best present solution) and the drone (the trial solution) according to their fitness values.
(7) Use the crossover and mutation mechanisms, and improve the new solution set according to the fitness values.
(8) Judge whether there is a better solution. If yes, keep the original optimal solution; if not, replace the optimal solution.
(9) Judge whether conditions for termination are met. If yes, stop the calculation; if not, discard all of the previous trial solutions, generate new trial solutions, and return to Step (4).

Stage 4: Performance evaluation

This paper tested the aforesaid process and methods. In previous research studies, four indexes were used to validate the classification effect, including the accuracy rate, precision, recall rate and F-measure. Precision is a measure of exactness, whereas recall is a measure of completeness. To be more specific, a precision rate is the fraction of retrieved instances that are relevant, while a recall rate is the fraction of relevant instances that are retrieved. The recall rate and accuracy rate were proposed by Salton [27]. A high recall rate means the classifier can learn the actual classification results effectively, and a high accuracy rate means the

classification results have high correctness.

A high recall rate and accuracy rate are achieved in ideal conditions; however, in reality it is difficult to achieve them concurrently. In general cases, the accuracy rate will therefore be increased as much as possible at the recall rate within a fixed range to attain the goal. The F-measure proposed by van Rijsbergen [28] evaluates the performance of the classifier. The F-measure combines the recall rate with the accuracy rate to express the efficiency of the overall classifier. The larger the F-measure value is, the higher the accuracy after classification will be. Three indexes were used to validate the classification effect of this paper, including the precision, recall rate and F-measure.

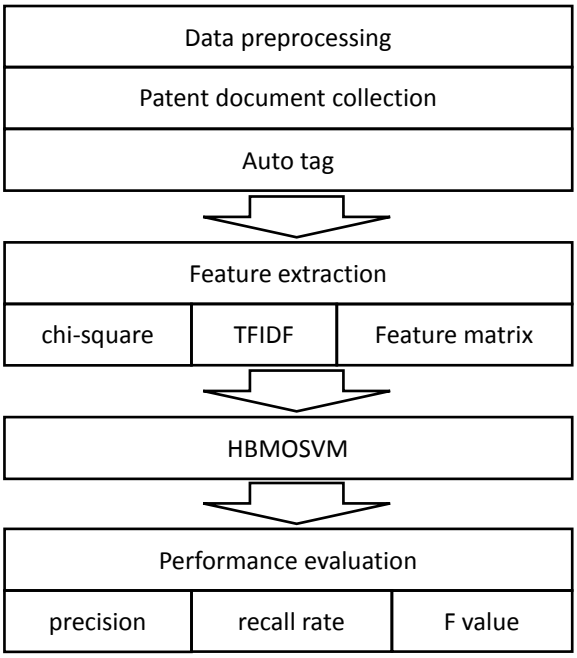

Figure 6. Case analysis flow chart.

Experimental Results and Analysis

Most of the SVM parameters are determined by correction methods. Generally speaking, there is no absolute standard for the parameters; however, they are very important for the SVM classification process. Therefore, this study placed emphasis on the selection of parameters, and used the extent of solution and the disturbance mechanism in the swarm algorithm to seek the optimal parameters. In addition, regarding patent classification, the extraction of the feature number not only influences the effect of classification but also influences the efficiency of classification. First, the chi-square statistics and traditional SVM were used to find out the appropriate number of features, and then the HBMOSVM of this study was used to seek the optimization parameters, so as to improve the accuracy of classification.

Experimental Environment

This study used improved classifiers for patent document classification, based on TM preprocessing, text feature extraction and the patent algorithm. The on-line auto tag system developed by Academia Sinica was used in TM preprocessing [29], and Matlab 7.6 was used for the TFIDF TF and IDF algorithms. The same development environment was used for chi-square statistics in the text feature extraction, and the improved hybrid algorithm Libscm 2.91 [30], developed by Dr. Chi-wen Lin of National Taiwan University, was used to validate the cases of this study.

Test Data Collection

The empirical data in this study were Taiwan patent data provided by the Intellectual Property Office of the Ministry of Economic Affairs. There were approximately 860,000 pieces of patent data for the period of January 1998 and December 2009. As the classes of the various patent data needed to be defined prior to the training process of automatic classification, this paper used CMP as the main target data set. A total of 210 articles were downloaded, and the patent titles and patent summarizations were collected. The following figure shows the IPC classification code based on a hierarchical classification structure, which became the output class of the SVM in classification. In Figure7, C09G stands for “Polishing compositions other than French polish; Ski waxes”, B24B stands for “Machines, devices, or processes for grinding or polishing; Dressing or conditioning of abrading surfaces; Feeding of grinding, polishing, or lapping agents”, and H01L stands for “Semiconductor devices; Electric solid state devices not otherwise provided for.”

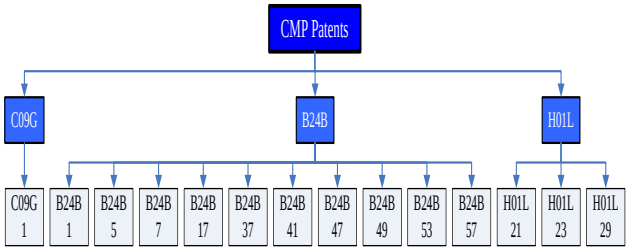

Figure 7. CMP output class structure diagram.

Training Recognition of HBMOSVM

In this study, 210 patent documents were used as the test documents. The text mining technology preprocessing reduced the feature dimensions of the input matrix by 5, 15, and 25, and the three feature number data sets were compared. The GASVM and

HBMOSVM of this study were used in 20 experiments to analyze the influence of the differences in the dimensions and algorithms on the classification accuracy. The average values of the 20 tests are listed in Table 2, and the evaluation indexes after the tests are listed in Table 3 and Table 4. The Mann-Whitney U test (Table 3) indicated that under three different feature thresholds, the classification results for HBMOSVM were significantly different from that of GASVM, and the accuracy rate approached a high classification level of accuracy and precision. However, as the feature number increased, although the classification indexes could be improved slightly, the computing time was prolonged greatly. Therefore, this study chose the data of feature number 15 as the imported high dimensional sparse matrix. Figure 8 shows the trend of the F-measure. It was found that both the F-measure and accuracy increased with the feature number, and that the classification evaluation index F-measure of HBMOSVM was better than that of GASVM.

Table 2. Training recognition results.

Feature dimensions	GASVM(%)	HBMOSVM(%)
Chi(5)	69.38	79.64
Chi (15)	83.90	92.19
Chi (25)	91.62	94.24

Table 3. Efficiency test for GASVM and HBMOSVM at different feature numbers.

Mann-Whitney U Test			
Feature dimensions	Chi(5)	Chi(15)	Chi(25)
Mann-Whitney U	42.500	34.500	150.000
Z	-4.269	-4.691	-2.043
P-value	.000	.000	.041

Table 4. Training evaluation index results.

Algorithm	Feature dimensions	Evaluation index	B24B	C09G	H01L
GA SVM	**Chi(5)**	Precision	0.7903	0.8571	0.5333
		Recall	0.7101	1.000	0.8889
		F-measure	**0.7481**	**0.9231**	**0.6667**
	Chi (15)	Precision	0.8667	0.9333	0.9556
		Recall	0.8843	1.000	0.9149
		F-measure	**0.8754**	**0.9655**	**0.9231**
	Chi (25)	Precision	0.9200	1.000	1.000
		Recall	0.9200	1.000	1.000
		F-measure	**0.9200**	**1.000**	**1.000**
HBMO SVM	**Chi (5)**	Precision	0.8467	1.000	0.800
		Recall	0.8038	1.000	0.9730
		F-measure	**0.8247**	**1.000**	**0.8780**
	Chi (15)	Precision	0.8933	1.000	1.000
		Recall	0.8993	1.000	0.9783
		F-measure	**0.8963**	**1.000**	**0.9890**
	Chi (25)	Precision	0.9200	1.000	1.000
		Recall	0.9200	1.000	1.000
		F-measure	**0.9200**	**1.000**	**1.000**

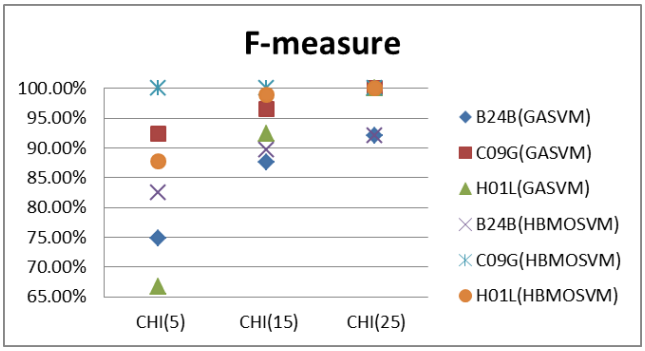

Figure 8. F-measure trend map.

Prediction Recognition Results of HBMOSVM

The forecast ability of the HBMOSVM model was tested according to the ratio of training documents to test documents. This study used 7:3 and 6:4 sample ratios for the experiments, and used the high dimensional input matrix of the data for feature numbers 15 and 25 for the test. Table 5 and Table 6 show the test results of different ratios. The data showed that the classification accuracy rate of the data samples in different ratios tested by HBMOSVM was higher than that of GASVM. Table 7 shows the differences in the accuracy of the two algorithms. The data show that in four cases, the accuracy of HBMOSVM is more significant than that of GASVM (P-value<0.05). Table 8 and Table 9 show the detailed results of the classification of test documents. In terms of the F-measure evaluation index, Figure 9 and Figure 10 show that the evaluation index of HBMOSVM was higher than that of GASVM. A cross-over experiment was carried out according to two different influencing factors, prior to the analysis of the patent document's forecast ability. It was found that when the HBMOSVM algorithm carried out the classification, the performance indicator was better than the classification performance of the GASVM algorithm, and that it would not change when other factors changed. In other words, this algorithm was stable. The classification accuracy rate and the F-measure for the performance evaluation were the results of the performance, meaning it is feasible to use the HBMOSVM algorithm as a prediction model for classification.

Table 5. Test results of two algorithms when the sample ratio is 7:3.

Algorithm	GASVM		HBMOSVM	
Feature dimensions	Chi(15)	Chi(25)	Chi(15)	Chi(25)
Accuracy rate(%)	**74.37**	**75.71**	**78.17**	**81.11**

Table 6. Test results of two algorithms when the sample ratio is 6:4.

Algorithm	GASVM		HBMOSVM	
Feature dimensions	Chi(15)	Chi(25)	Chi(15)	Chi(25)
Accuracy rate(%)	**75.60**	**76.01**	**81.25**	**83.15**

Table 7. Efficiency test for two algorithms under ratio factors.

Mann-Whitney U Test				
Sample Ratio	7:3		6:4	
Feature Dimensions	Chi(15)	Chi(25)	Chi(15)	Chi(25)
Mann-Whitney U	107.5	80.0	75.5	30
Z	-2.525	-3.412	-3.385	-4.619
P-value	.012	.001	.001	.000

Table 8. Classification indexes of two algorithms when the sample ratio is 7:3.

Algorithm	Feature dimensions	Evaluation index	B24B	C09G	H01L
GA SVM	Chi (15)	Precision	0.9535	0.3750	0.5000
		Recall	0.8039	1.0000	0.6667
		F-measure	**0.8723**	**0.5455**	**0.5714**
	Chi (25)	Precision	1.0000	0.4000	0.3846
		Recall	0.8036	1.0000	1.0000
		F-measure	**0.8911**	**0.5714**	**0.5556**
HBMO SVM	Chi (15)	Precision	1.0000	0.5000	0.4667
		Recall	0.8148	1.0000	1.0000
		F-measure	**0.8980**	**0.6667**	**0.6364**
	Chi (25)	Precision	0.9800	0.25	0.2222
		Recall	0.8448	1	0.5
		F-measure	**0.9074**	**0.4**	**0.3076**

Table 9. Classification indexes of two algorithms when the sample ratio is 6:4.

Algorithm	Feature dimensions	Evaluation index	B24B	C09G	H01L
GA SVM	Chi (15)	Precision	0.9623	0.3333	0.7273
		Recall	0.8226	1.0000	0.8421
		F-measure	**0.8870**	**0.5000**	**0.7805**
	Chi (25)	Precision	1.0000	0.5000	0.4000
		Recall	0.7945	1.0000	1.0000
		F-measure	**0.8855**	**0.6667**	**0.5714**
HBMO SVM	Chi (15)	Precision	0.9839	0.4286	0.5333
		Recall	0.8592	1.0000	0.8000
		F-measure	**0.9173**	**0.6000**	**0.6400**
	Chi (25)	Precision	0.9836	0.4286	0.5625
		Recall	0.8451	1.0000	0.9000
		F-measure	**0.9091**	**0.6000**	**0.6923**

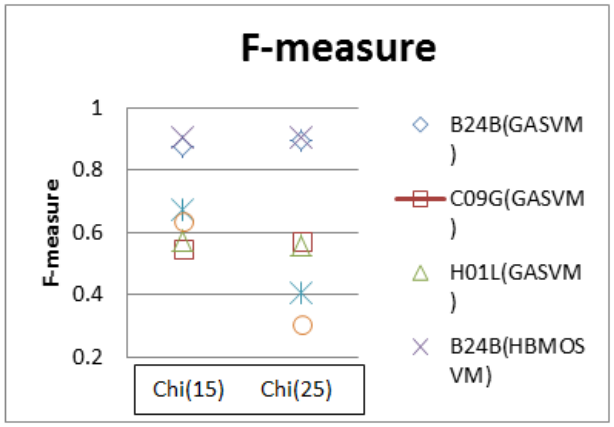

Figure 9. F-measure trend map when the sample ratio is 7:3.

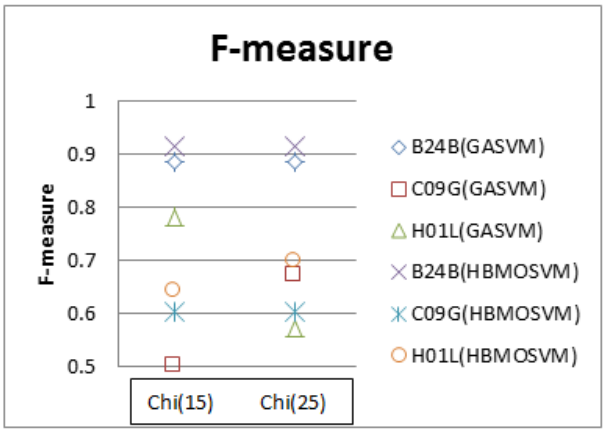

Figure 10. F-measure trend map when the sample ratio is 6:4.

Conclusions

The SVM is one of the best tools in the TM domain. In the patent document classification procedure, text mining technology is used first to integrate documents, and then they are converted into the vector data of a word matrix by the kernel function and projected into a high dimensional feature space. Finally, the target classifier is used to look for the optimum tangent plane of various classes in the feature space, as the basis for classification. The results of this study showed that HBMOSVM had a high accuracy classification result and performance compared with traditional SVM, and that it could shorten the time spent on the search process for optimum parameters.

It was found in the proof procedure of patent document classification that the factors influencing the patent document classification results were the feature number and the number of training data. The traditionally common feature extraction methods are document frequency, information acquisition, chi-square statistics, term strengthening and the document frequency-entropy approach. This study used the chi-square statistics for feature extraction to select highly correlated keywords as the input matrix, and a total of

6,460 keywords were obtained. However, the feature number was selected by considering the fluency of the calculation. The test results showed that without influencing classification accuracy or system computation complexity, the accuracy of the two feature input matrices of feature numbers 15 and 25 was as high as 92% after classification by the HBMOSVM proposed in this study. Therefore, these two feature numbers were used for the prediction test simulation in Stage 2.

During the second stage, in order to evaluate the effect of different amounts of training data on the classification capacity of the model, the number of training data (60% and 70% of the number of samples) was added for data analysis. The result showed that the classification accuracy of the prediction model which used 126(=210*0.6) training samples was higher than the other prediction model which used 147(=210*0.7) training samples; however, it was not very different. In terms of the HBMOSVM algorithm, the classification performance of the model which included 60% training data was better than the performance of the model with 70% training data, meaning the effect of the data ratio on the accuracy was lower than that of the feature number. As long as the ratio of training data was kept in a fixed range, good performance classification would occur.

The selection of the algorithm is still the most significant factor influencing the patent document classification effect. According to the case validation results, the HBMOSVM could result in better patent documentation accuracy and better F-measure performance as an evaluation index than GASVM under different factors. The influencing factor was that the algorithm remedied the deficiency in selecting the SVM parameters. The algorithm was improved, the performance indicator was increased, and the complex computing process was reduced. Therefore, the HBMOSVM proposed in this study had good classification accuracy and stability for a reference data set, and it could solve and improve the rising patent document classification performance problem. For future study, the proposed HBMOSVM algorithm can be applied in management decision making concerning market segmentation and in customer relationship management.

References

[1] M. Fattori, G. Pedrazzi, and R. Turra, "Text mining applied to patent mapping: A practical business case," *World Patent Information,* vol. 25, no. 4, pp. 335-342, 2003.
doi: 10.1016/S0172-2190(03)00113-3

[2] C. Cortes and V. Vapnik, "Support-vector networks," *Machines Learning,* vol. 20, no. 3, pp. 273-297, 1995.
doi: 10.1023/a:1022627411411

[3] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Communication of the ACM,* vol. 18, no. 11, pp. 613-620, 1975.
doi: 10.1145/361219.361220

[4] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," The 5th Annual ACM Workshop on COLT, Pittsburgh, Pennsylvania, USA, pp. 144-152, 1992.
doi: 10.1145/130385.130401

[5] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," in *Advances in Neural Information Processing Systems 9 (NIPS)*, MIT Press, 1996, pp. 155-161.

[6] K. Z. Mao, "Feature subset selection for support vector machines through discriminative function pruning analysis," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics,* vol. 34, no. 1, pp. 60-67, 2004.
doi: 10.1109/TSMCB.2002.805808

[7] V. N. Vapnik, *Statistical learning theory*, Wiley, 1998.

[8] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowledge Discovery*, vol. 2, no. 2, pp. 121-167, 1998.
doi: 10.1023/a:1009715923555

[9] B. Schölkopf, *Support vector learning*. Munich, Germany: Oldenburg-Verlag, 1997.

[10] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks,* vol. 10, no. 5, pp. 988-999, 1999.
doi: 10.1109/72.788640

[11] H. A. Abbass, "Marriage in honey bees optimization (mbo): A haplometrosis polygynous swarming approach," in *the Congress on Evolutionary Computation,*, Seoul, Korea, 2001, vol. 1, pp. 207-214.
doi: 10.1109/CEC.2001.934391

[12] H.A. Abbass, "A monogenous MBO approach to satisfiability", in *the International Conference on Computational Intelligence for Modeling Control and Automation*, CIMCA, Las Vegas, NV, USA, 2001.

[13] O. Haddad, A. Afshar, and M. Mariño, "Honey-bees mating optimization (HBMO) algorithm: A new heuristic approach for water resources optimization," *Water Resources Management,* vol. 20, no. 5, pp. 661-680, 2006.
doi: 10.1007/s11269-005-9001-3

[14] A. Afshar, O. Bozorg Haddad, M. A. Mariño, and B. J. Adams, "Honey-bee mating optimization (HBMO) algorithm for optimal reservoir operation," *Journal of the Franklin Institute,* vol. 344, no. 5, pp. 452-462, 2007.
doi: 10.1016/j.jfranklin.2006.06.001

[15] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science,* vol. 220, no. 4598, pp. 671-680, 1983.
doi: 10.1126/science.220.4598.671

[16] M. Marinaki, Y. Marinakis, and C. Zopounidis, "Honey Bees Mating Optimization algorithm for financial classification problems," *Applied Soft Computing,* vol. 10, no. 3, pp. 806-812, 2010.
doi:10.1016/j.asoc.2009.09.010

[17] T. Niknam, "Application of honey-bee mating optimization on state estimation of a power distribution system including distributed generators," *Journal of Zhejiang University SCIENCE A,* vol. 9, no. 12, pp. 1753-1764, 2008.
doi: 10.1631/jzus.A0820047

[18] O. Bozorg Haddad, M. Mirmomeni, M. Zarezadeh Mehrizi, and M. Mariño, "Finding the shortest path with honey-bee mating optimization algorithm in project management problems with constrained/unconstrained resources," *Computational Optimization and Applications,* vol. 47, no. 1, pp. 97-128, 2010.
doi: 10.1007/s10589-008-9210-9

[19] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurasamy, *Advances in knowledge discovery and data mining*. Menlo Park, California: AAAI Press : MIT Press, 1996.

[20] R. J. Mooney and U. Y. Nahm, "Text mining with information extraction," in *the 4th International MIDP Colloquium*, Van Schaik Pub., South Africa, 2005, pp. 141-160.

[21] S. Lo, "Web service quality control based on text mining using support vector machine," *Expert Systems with Application,* vol. 34, no. 1, pp. 603-610, 2008.
doi: 10.1016/j.eswa.2006.09.026

[22] C.-W. Chang, C.-T. Lin, and L.-Q. Wang, "Mining the text information to optimizing the customer relationship management," *Expert Systems with Applications,* vol. 36, no. 2, Part 1, pp. 1433-1443, 2009.
doi: 10.1016/j.eswa.2007.11.027

[23] A. Pons-Porrata, R. Berlanga-Llavori, and J. Ruiz-Shulcloper, "Topic discovery based on text mining techniques," *Information Processing & Management,* vol. 43, no. 3, pp. 752-768, 2007.
doi: 10.1016/j.ipm.2006.06.001

[24] D. Delen and M. D. Crossland, "Seeding the survey and analysis of research literature with text mining," *Expert Systems with Applications,* vol. 34, no. 3, pp. 1707-1720, 2008.
doi: 10.1016/j.eswa.2007.01.035

[25] C. Hung, Y.-L. Chi, and T.-Y. Chen, "An attentive self-organizing neural model for text mining," *Expert Syst. Appl.,* vol. 36, no. 3, pp. 7064-7071, 2009.
doi: 10.1016/j.eswa.2008.08.037

[26] H. T. Ng, W. B. Goh, and K. L. Low, "Feature selection, perception learning, and a usability case study for text categorization," in *the 20th annual international ACM-SIGIR conference on Research and development in information retrieval*, Philadelphia, Pennsylvania, USA, 1997, pp. 67-73.
doi: 10.1145/258525.258537

[27] G. Salton, *Automatic text processing : The transformation, analysis, and retrieval of information by computer*. New York: Addison-Wesley, 1989.

[28] C. J. Van Rijsbergen, *Information retrieval*. London, England: Butterworths, 1979.

[29] *The on-line auto tag system*, [Online].
Available: http://ckipsvr.iis.sinica.edu.tw/

[30] *LIBSVM*, [Online].
Available: www.csie.ntu.edu.tw/~cjlin/libsvm/