Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications xxx (2008) xxx–xxx
Expert Systems with Applications
www.elsevier.com/locate/eswa

Construction of supervised and unsupervised learning systems for multilingual text categorization

Chung-Hong Lee [a,*], Hsin-Chang Yang [b]

[a] Department of Electrical Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan
[b] Department of Information Management, National University of Kaohsiung, Kaohsiung, Taiwan

Abstract

Due to the availability of a huge amount of textual data from a variety of sources, users of internationally distributed information regions need effective methods and tools that enable them to discover, retrieve and categorize relevant information, in whatever language and form it may have been stored. This drives a convergence of numerous interests from diverse research communities focusing on the issues related to multilingual text categorization. In this work, we implemented and measured the performance of the leading supervised and unsupervised approaches for multilingual text categorization. We selected support vector machines (SVM) as representative of supervised techniques as well as latent semantic indexing (LSI) and self-organizing maps (SOM) techniques as our selective ones of unsupervised methods for system implementation. The preliminary results show that our platform models including both supervised and unsupervised learning methods have the potentials for multilingual text categorization.

Keywords: Text categorization; Text Mining; Machine learning; Supervised learning; Unsupervised learning

1. Introduction

The availability of a huge amount of textual data from a variety of sources leads to the explosion and overloads of information. Furthermore, the most valuable information is encoded in pages which are written in various native languages, but are relevant even to non-native speakers. The process of accessing and categorizing all these raw data, heterogeneous for language used, and transforming them into information is therefore inextricably linked to the concepts of textual analysis and synthesis, hinging greatly on the ability to master the problems of multilingualism. Through automatic categorization of multilingual texts, users can get an overview of great volumes of textual data having a highly readable grid, which helps them organize relevant multilingual documents and find all related information.

As text categorization systems become an essential part of organizational knowledge management systems, adaptability to variations in textual languages and dynamics of application scenarios becomes increasingly important. It is therefore imperative to move towards adaptive categorization systems that selectively employ appropriate categorization method(s) by first analyzing the available data sources, or directly developing multilingual text categorization systems based upon a unified algorithmic platform. For the first case, text categorization systems should support different techniques and apply the most appropriate method(s) that suits the data characteristics of the problem at hand. However, in order to build adaptive text categorization systems, one must first understand the performance characteristics of the categorization methods in a systematic manner. Development of multilingual text categorization systems seems to be a comparatively feasible and effective way to meet the requirements of organizing multilingual digital documents.

* Corresponding author. Tel.: +886 7 3814526x5581; fax: +886 7 3921073.
E-mail addresses: leechung@mail.ee.kuas.edu.tw (C.-H. Lee), yanghc@nuk.edu.tw (H.-C. Yang).

0957-4174/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2007.12.052

There are mainly two machine learning approaches to enhance this task: supervised learning techniques, where pre-defined category labels are assigned to documents based on the likelihood suggested by a training set of labeled documents; and unsupervised learning approaches, where there is no need for human intervention or labeled documents at any point in the whole process. The original motivation of this work is based on the fact that, although various supervised machine learning techniques have been effectively applied to text categorization, it still confronted some challenging issues. Due to the rapidly increasing amounts of online documents, the dynamic nature of most text databases makes it difficult to pre-define the categories. One significant difficulty with these current algorithms, and the issue addressed by this paper, is that supervised techniques require the extra effort to predefine the categories and to assign category labels to the documents in the training set. In addition, they normally need a large number of labeled training examples to learn accurately. Labeling must often be done by a person; this is a painfully time-consuming process. For example, for classification of collected news articles, systems that filter or pre-sort articles and present only the ones the user finds interesting are highly desirable, and are of great commercial interest today. A past paper in the literature reported that after a person read and labeled about 1000 articles, a learned classifier achieved a precision of about 50% when making predictions for only the top 10% of documents about which it was most confident. Most users of a practical system, however, would not have the patience to label a thousand articles-especially to obtain only this level of precision. One would obviously prefer algorithms that can provide accurate classifications after hand-labeling only a few articles, rather than thousands. Furthermore, for a supervised categorization, different human experts may disagree when deciding under which category to categorize a given document. The need for large quantities of data to obtain high accuracy, the difficulty of obtaining labeled data, and the subjectivity in assigning documents to categories, raises an important question: what other approaches can reduce the need for labeled data? This has led us to consider that perhaps by nature text organization should be an unsupervised task rather than a supervised one. Given that each categorization method has its strengths and limitations and that real world problems do not always satisfy the assumptions of a particular method, one approach is to apply all appropriate methods and select the one that provides the best solution. The view is taken, therefore, our goal in this study is to implement categorization systems using different techniques and measure the performance of the selected supervised and unsupervised techniques for text categorization. We hope that unsupervised approaches can be used to give feedback to the human indexers to enhance the task of pre-defining categories and preparing a labeled training set. This study will possibly form the basis for developing a hybrid approach of supervised and unsupervised paradigm to the domain of text categorization by also considering the limitations mentioned above. In this work we employed and measured the performance of the leading unsupervised and supervised approaches for multilingual text categorization by using various standard document corpora. We selected self-organizing maps (SOM) and latent semantic indexing (LSI) techniques as representatives of unsupervised methods as well as support vector machines (SVM) (Joachims, 1998; Vapnik, 1995; Vapnik, 1999) as a representative of supervised techniques for implementation, respectively.

The rest of this paper is organized as follows. In Section 2, we describe the concepts of supervised and unsupervised learning techniques associated with the implementations of multilingual text categorizations. Section 3 presents experiments with an unsupervised technique for system implementation. The experiments with a supervised technique for system implementation are described in Section 4. In Section 5, we present and discuss the experimental results. In Section 6, we discuss the performance measures, based on the comparative study of the experimented systems. Section 7 introduces the related work reported in the literature. Finally, we reach our conclusion in Section 8.

2. Preliminary

Multilingual text categorization (MTC) is a relatively new research topic, about which not much previous work in the literature appears to be available. Still, it concerns a practical problem, which is increasingly felt in some application fields, such as the documentation departments of international organizations as they come to rely on automatic text classification. It is also manifest in many news sites on the web, which rely on a quick classification of multinational news information. One major difficulty of multilingual text classification is the complexities of languages in the text contents, along with their high dimensional nature of document data sets. In this work, we attempt to tackle such a problem domain with different types of learning techniques, including supervised and unsupervised learning methods. In the literature, the document categorization using machine learning techniques normally consider the supervised methods to carry out the tasks. Due to that each categorization method has its strengths and drawbacks and that real world problems do not always satisfy the assumptions of a particular method, our approach is to apply all appropriate methods and review their perspective framework processes for obtaining the best solutions and tradeoffs. Through implementing supervised and unsupervised learning approaches, we aim to explore and demonstrate the potentials of both learning techniques in the applications of multilingual text categorization.

2.1. Supervised learning approaches

As mentioned above, automatic text categorization is normally performed by supervised learning techniques,

Please cite this article in press as: Lee, C. -H. , & Yang, H.-C., Construction of supervised and unsupervised learning systems ..., *Expert Systems with Applications* (2008), doi:10.1016/j.eswa.2007.12.052

where pre-defined category labels are assigned to documents based on the likelihood suggested by a training set of labeled documents. Many learning algorithms such as k-nearest neighbor (k-NN) (Masand, Linoff, & Waltz, 1992; Yang & Liu, 1999; Yang & Pedersen, 1997), support vector machines (SVM) (Joachims, 1998), neural networks (Ng, Goh, & Low, 1997; Wiener, Pedersen, & Weigend, 1995), linear least squares fit (LLSF) (Yang & Liu, 1999), and naïve Bayes (NB) (Koller et al., 1997; McCallum & Nigam, 1998) have been applied to text classification. A comparison of these techniques is addressed by Yang (1999) and Yang and Liu (1999). They conclude that all these approaches perform comparably when each category contains over 300 documents. However, when the number of positive training documents per category is less than 10, SVM, k-NN and LLSF outperform significantly neural networks and NB. These techniques did provide state-of-the-art learning approaches to represent a viable and well-performing solution for monolingual categorization problems. Unfortunately, little attention has been paid for practical applications which need applying text categorization approaches to multilingual information sources in the real-world scenario. To pursue a salient supervised way for MTC and a comparable technique competed with unsupervised methods, in this work we developed a supervised-classifier technique to enable multilingual documents be effectively categorized, by means of utilizing a number of selected multilingual corpora to train the classifiers based on the *support vector machines* model.

2.2. Unsupervised learning approaches

In unsupervised text categorization, we have unlabelled collection of documents in multiple languages. The aim is to cluster the documents without additional knowledge or intervention such that documents within a cluster are similar than documents between clusters. Self-organizing maps (SOM) techniques provide document clustering and word clustering methods to group similar texts. In our previous projects, we have successfully demonstrated the potentials of the SOM methods in dealing with organizing complex texts by applying them to several text mining applications, including multilingual text mining, automatic generation of web directories, Chinese text categorization, automatic construction of hypertexts and image semantics discovery (Lee & Yang, 2003; Yang & Lee, 2004; Yang & Lee, 2005a, 2005b; Yang et al., 2008). The SOM models have been proven to be powerful unsupervised methods to discover similarities among documents for text categorization. Latent Semantic Indexing (LSI) constitutes a paradigm that groups words into 'concepts' based on their co-occurrences in a given dataset (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Littman, Dumais, & Landauer, 1998). LSI then allows for text clustering or classification taking into account these 'concepts'. Its biggest advantage is that a thesaurus (e.g. WordNet) is not needed, but the largest disadvantage of LSI is that the notion of 'concept' that LSI introduces cannot easily be explained to a common user. In this work, we adopt the SOM and LSI models to the implementation of unsupervised learning platform. The judgments for the technique selection for the MTC system implementation is described in the following section.

2.3. Considerations for selecting unsupervised and supervised techniques to implementation

Most traditional text retrieval tasks are performed based upon vector space model (VSM) (Salton & McGill, 1983) by using flat feature representation of documents to reflect the characteristics of the sparse feature space in document representation. In the VSM model, the frequency of each term of the vocabulary is counted for a given document. A term weight vector is then constructed for a document using this "term frequency" together with "document frequency". However, the length of VSM feature is normally too long so that it becomes practically not useful for large-scale application. For tackling such difficulties, several vector based approaches have been used to compress large histogram vector into low dimensional feature. Among these methods, the unsupervised learning approaches such as LSI and SOM based models, are shown to be effective in encoding the semantics of documents without the penalty of losing large document information. It is beneficial for fulfilling accurate text categorization tasks in the vector space. In addition to unsupervised methods, some supervised learning methods such as SVM models are capable of effectively processing high dimensional feature vectors for the representation of texts in input space. In the vector space of text feature, the supervised SVM with kernel functions and unsupervised LSI and SOM methods imply different merits in text categorization respectively. As a result, in this work we select them (i.e. LSI, SOM and SVM models) as representatives for a study on implementation of multilingual text categorization systems.

3. Implementation of unsupervised methods for MTC

3.1. Document preprocessing

Our approach begins with a standard practice in information retrieval (IR) to encode documents with vectors, in which each component corresponds to a different word, and the value of the component reflects the frequency of word occurrence in the document. In other words, it creates a multidimensional vector space model for a given set of text documents through extracting unique content-bearing items, including words or phrases, and then to treat those items as features representing each document as a vector. We may regard this vector space model as a word-to-document feature-appearance matrix where rows are the features and columns are document vectors. For a Chinese corpus, the document preprocessing is relatively complicated. Since a Chinese sentence is composed of

Please cite this article in press as: Lee, C. -H. , & Yang, H.-C., Construction of supervised and unsupervised learning systems ..., *Expert Systems with Applications* (2008), doi:10.1016/j.eswa.2007.12.052

characters without boundaries, segmentation is indispensable. We employ a dictionary, some morphological rules and an ambiguity resolution mechanism for segmentation. In addition, we also extract named organizations, people, and locations, along with date/time expressions and monetary and percentage expressions. The rest of the process is the same as that of text indexing in an English corpus. The encoded document vectors were then used to train the text classifiers and using unsupervised and supervised learning techniques respectively, for further studies.

In unsupervised text categorization techniques, we have unlabeled collection of documents. The aim is to categorize the documents without additional knowledge or intervention such that documents within a category (or a cluster).

It is worth mentioning that during the process of text categorization, features are converted to a multidimensional vector space and each feature-dimension is assumed to be independent of other features. This assumption simplifies the process of classification but will also depilate the connections among different features; therefore, it may result in losing semantic information. However, it has generally been accepted that to a great extent the unordered combinations of features are still capable of representing the content information of a document. This concept is essential to allow further applications such as LSI based methods in representing texts for classification. The development process of the selected unsupervised systems (i.e. SOM and LSI based models) is illustrated in Fig. 1.

3.2. Self-organizing map methods

The *self-organizing map* (*SOM*) (Kohonen, 1982, 1995) is one of the major unsupervised artificial neural network models. It basically provides a way for cluster analysis by producing a mapping of high dimensional input vectors onto a two-dimensional output space while preserving topological relations as faithfully as possible. After appropriate training iterations, the similar input items are grouped spatially close to one another. As such, the resulting map is capable of performing the clustering task in a completely unsupervised fashion. In this work we employ the SOM method to produce two maps for text categorization, namely the *word cluster map (WCM)* and the *document cluster map (DCM)*. The construction of the SOM model for MTC implementation is shown in Fig. 2. The concepts of WCM and DCM are described in the following subsection.

3.2.1. The SOM based text categorization algorithm

The word cluster map that is employed for document encoding is produced according to word similarities, measured by the similarity of the co-occurrence of the words. Conceptually related words tend to fall into the same or neighboring map nodes. By means of the SOM algorithm, word clusters can be ordered and organized as nodes on the map. Let $\mathbf{x}_i \in \Re^N$, $1 \leqslant i \leqslant M$, be the feature vector of the ith document in the corpus, where N is the number of indexed terms and M is the number of documents. We used these vectors as the training inputs to the map. The map consists of a regular grid of processing units called *neurons*. Each neuron in the map has N synapses. Let $\mathbf{w}_j = \{w_{jn} | 1 \leqslant n \leqslant N\}$, $1 \leqslant j \leqslant J$, be the synaptic weight vector of the jth neuron in the map, where J is the number of neurons on the map. We trained the map by the SOM algorithm:

Step 1. Randomly select a training vector $\mathbf{x}_i$ from the corpus.

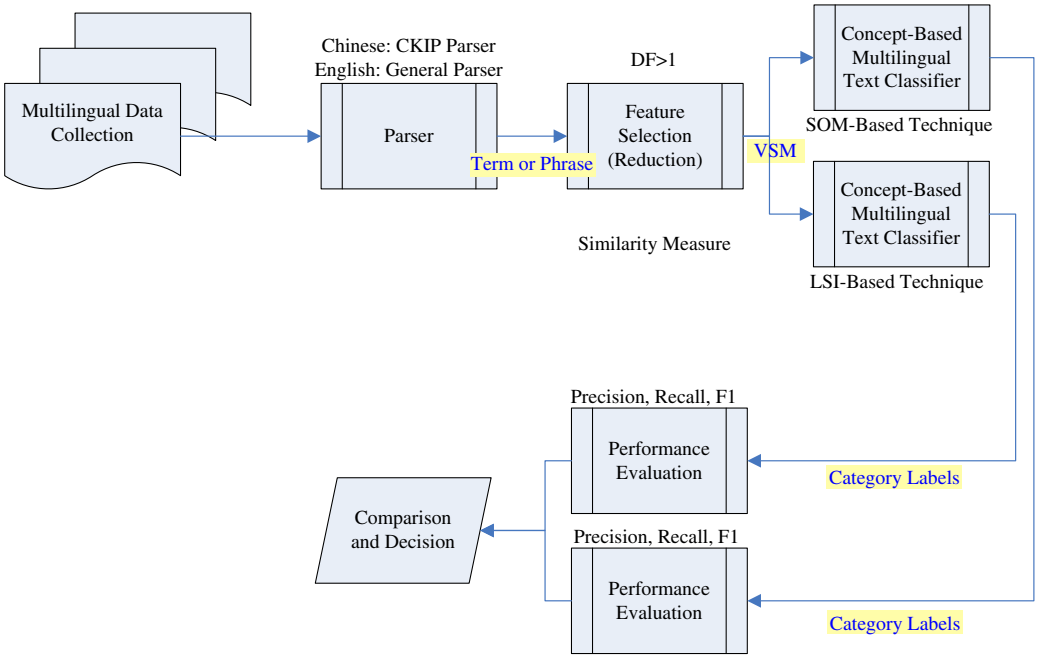

Fig. 1. Development process of unsupervised systems (i.e. SOM and LSI techniques).

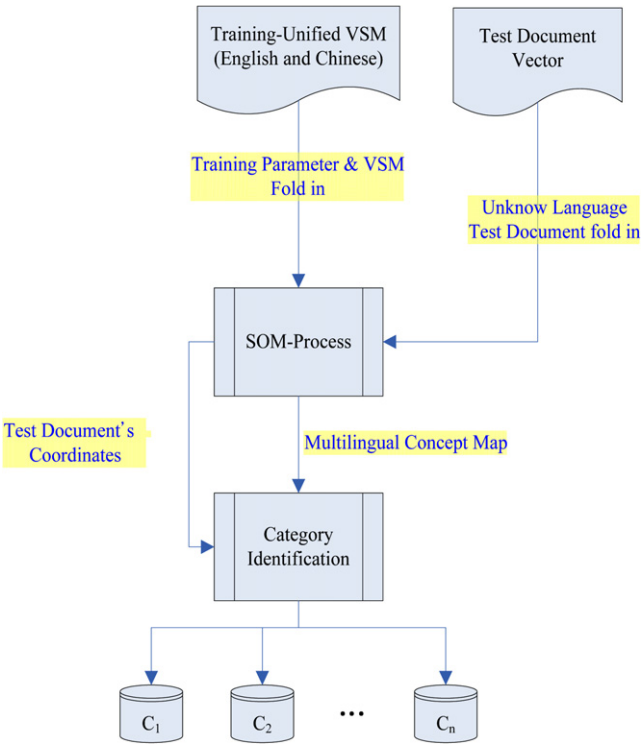

Fig. 2. The self-organizing map model for MTC.

Step 2. Find the neuron j with synaptic weights $\mathbf{w}_j$ which is closest to $\mathbf{x}_i$, i.e.

$$\|\mathbf{x}_i - \mathbf{w}_j\| = \min_k \|\mathbf{x}_i - \mathbf{w}_k\|. \quad (1)$$

Step 3. For every neuron l in the neighbor of node j, update its synaptic weights by

$$\mathbf{w}_l^{\text{new}} = \mathbf{w}_l^{\text{old}} + \alpha(t)(\mathbf{x}_i - \mathbf{w}_l^{\text{old}}), \quad (2)$$

where $\alpha(t)$ is the training gain at time stamp t.

Step 4. Increase time stamp t. If t reaches the preset maximum training time T, halt the training process; otherwise decrease $\alpha(t)$ and the neighborhood size, and go to Step 1.

The training process stops after time T which is sufficiently large that every feature vector may be selected as training input for certain times. The training gain and neighborhood size both decrease when t increases.

After the training process, the map forms a word cluster map (WCM) by labeling each neuron with certain words. For the nth word in the corpus we construct an N-dimensional vector $\mathbf{v}_n$ in which only the nth element is non-zero. To label the neurons, we present each $\mathbf{v}_n$ to the map and find the best matching neuron. Since the number of neurons is generally much smaller than the number of words, each neuron in the map may have multiple labels. We may say that a neuron forms a word cluster because the closely related words will map to the same neuron. The word cluster map autonomously clusters words according to their similarity of co-occurrence. Words that tend to be found in the same document will be mapped to close neurons in the map. For example, the Chinese words for "neural" and "network" often occur simultaneously in a document. They will map to the same neuron, or neighboring neurons, on the map. Words that do not occur in the same document will map to distant neurons on the map. Accordingly we can define the relationship between two words according to their corresponding neurons in the word cluster map, and the mining task will be performed based on such relationships. The trained map also forms a document cluster map (DCM) by labeling each neuron with certain documents. The document feature vectors $\mathbf{x}_i$ are presented to the map to label the neurons. Documents with similar keywords will map to the same or neighboring neurons. The similarity between two documents may be calculated by measuring the Euclidean distance between their mapped neurons in the map. Since the number of the neurons is much less than the number of the documents in the corpus, multiple documents may map to the same neuron. Thus, a neuron forms a document cluster. Besides, neighboring neurons represent document clusters of similar meaning, i.e. high keyword co-occurrence frequency.

After the training process, each neuron in the DCM and the WCM actually represents a document cluster and a word cluster, respectively. Such clusters can be considered as categories of the underlying corpus in text categorization task. After the categorization generation process, classification of documents into proper text categories can be achieved by our method.

3.3. Latent semantic indexing methods

Latent semantic indexing is a well-known technique in Information Retrieval, especially in dealing with polysemy and synonymy. LSI use SVD process to decompose the original term-document matrix into a lower dimension triplet. The triplet (the resulted matrices) is the approximation to original matrix and can capture the latent semantic relation between terms. The centroid of each class has been calculated in the decomposed SVD space. The similarity threshold of categorization is pre-defined for each centroid. Test documents with similarity measurement larger than the threshold will be labeled "Positive" (Relevant) or else would be labeled "Negative" (Non-Relevant). The LSI based techniques have been used as one of the major selections for the part of development of unsupervised methods for constructing the MTC system.

The system implementation by a developed unsupervised technique (i.e. a latent semantic indexing, LSI method) is described as follows. First, we established a cross-language vector space which would merge both English and Chinese documents into a single vector space model. We then apply the SVD (singular value decomposion) process onto the model to get the singular triplet which has already reduced the dimensions of the original term-document matrix.

Please cite this article in press as: Lee, C. -H. , & Yang, H.-C., Construction of supervised and unsupervised learning systems ..., *Expert Systems with Applications* (2008), doi:10.1016/j.eswa.2007.12.052

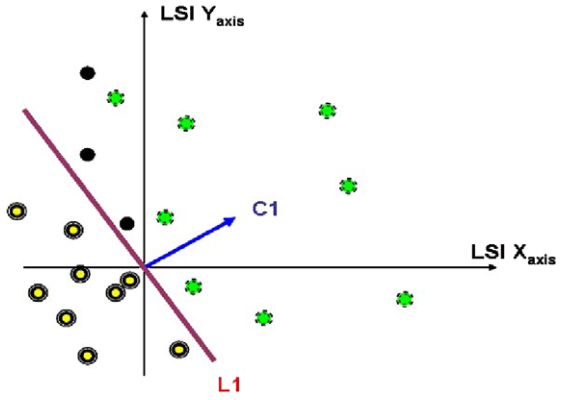

Fig. 3. C1 with coverage angle 90°.

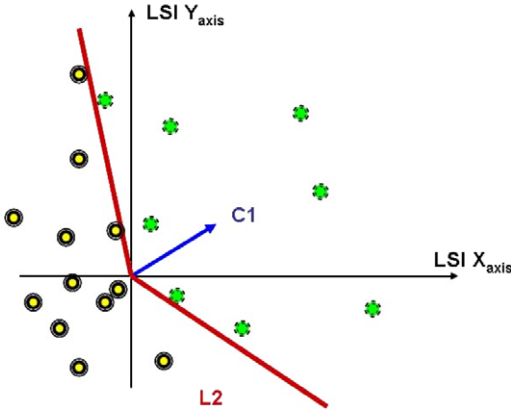

Fig. 4. C1 with coverage angle 75°.

Subsequently, we control two parameters called "centroid vector" and "coverage angle". Each class in the corpora has a corresponding centroid, we call them "class centroids". They are not physical but conceptual vectors which represent the documents,

$$C_i = \frac{1}{Ni} \sum \vec{d}_j \tag{3}$$

In Eq. (3), C_i is the class i centroid vector, N_i is the number of documents in class i, d_j is the document vector belonging to class i. Coverage angle is the angel between Centroid vector and threshold boundary.

If any document located outside the boundary, that is, the value of cosine (d, C_i) less than cosine (Coverage angel), It will be labeled "Neg" or "Non-Relevant" as indicated in Fig. 3.

In Fig. 3,[1] the green dots labeled "Pos" in class 1 are classified to "Pos". The yellow dots labeled "Neg" in class 1 are classified to "Neg". Both Yellow and Green dots are correctly classified, There are some classification error in Black dots.. The coverage angle is setting to 90° in the above figure, So that any document located in right side

[1] For interpretation of color representation in this figure the reader is referred to the web version of this article.

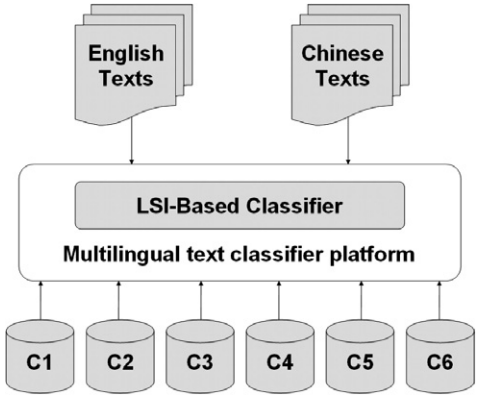

Fig. 5. The LSI-based classifier system.

of L1 is classify "Pos" ("Relevant") or classify "Neg" ("Non-Relevant"). Fig. 4 shows that there was no classification error if Coverage Angle was set to 75°. Fig. 5 shows the system framework for the LSI-based classifier system.

The platform illustrated in Fig. 5 indicates the resulting process of LSI-based MTC tasks. In the platform, we utilize the class centroid of training document in LSI-space to represent the concept of corresponding category, in which any document in LSI-space whose similarity with class centroid larger than pre-defined threshold (usually near to zero) would be classified to the category, and then performing MTC functions.

4. Implementation of supervised methods for MTC

In our previous work, we employed a supervised text mining technique based on support vector machines (SVMs) for training text classifiers in a combined platform (Lee & Yang, 2004; Lee & Yang, 2005; Lee, Yang, Hsu, Chen, & Hung, 2005). In this work we employed Chinese and English corpora as training sources for constructing the multi-classifier system. After training process, we used several unlabeled texts written in Chinese and English to evaluate the performance of categorization. The original idea of our framework is to assist people who try to categorize specific types of multilingual documents related to certain known fields. It is not designed to the general public who try to find information without much knowledge apriori most of the time. As a framework prototype, therefore, we used only six selected categories of texts in English and Chinese to train six support vector machines (SVM) classifiers, which are believed to be sufficient for defining a fundamental system model for testifying the theory of multilingual text categorization. This method works well if the original number of classes are limited, however, we allow the number of classes to be increased by the use of our system to a certain amount. We constructed SVM classifiers with a one-against-all (OAA) learning strategy to implement our multi-classifier system. Although there were still several learning strategies such as one-against-one (OAO) and DAGSVM methods applicable, the OAA

Please cite this article in press as: Lee, C. -H. , & Yang, H.-C., Construction of supervised and unsupervised learning systems ..., *Expert Systems with Applications* (2008), doi:10.1016/j.eswa.2007.12.052

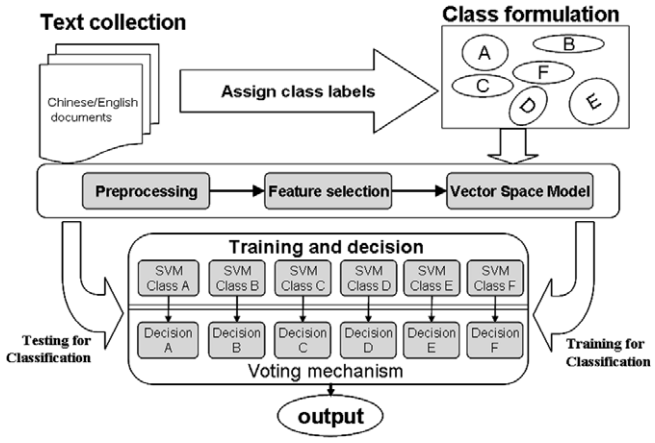

Fig. 6. System and process of the SVM-based MTC model.

technique selected for our implementation was based on the tradeoffs of the total costs of the amounts of classifiers and system performance. The OAA technique allows the deployment of fewer classifiers to achieve the functionalities of multi-class categorization and also obtain a reasonable performance in our application. The system implementation and experiments are described in the following subsections.

4.1. The formulation of SVM classifiers

The experimental process includes two phases. First, we are focusing on the generation of SVM classifiers by means of training of the experimental articles, covering Astronomy, Physics, Politics, Finance, Medicine and Arts categories. The classifiers were well developed based on the best results performed by the training and testing process mentioned above. Subsequently, we again take the pre-selected articles used in the training and testing process of SVM classifiers to formulate the classification results by the supervised method. The platform of the SVM based MTC model is illustrated in Fig. 6. The experimental results will be discussed in the later sections in detail.

5. Experimental results

5.1. Experiments with the SOM-based MTC system

We transform a document to a vector of word occurrence. After the self-organizing process, two documents will map to near neurons if they contain similar word occurrences. When different words are labeled on the same neuron or near neurons on the word cluster map, they tend to occur in a restricted set of documents. On the other hand, if two words seldom co-occur in any document, they should not be labeled on near neurons. This is because the neuron may be viewed as representing a virtual document containing those words labeled on it. Two words will be mapped to the same neuron if, and only if, they often co-occur in the same document, otherwise the virtual document may not contain these words simultaneously. Neighboring neurons in the word cluster map represent word clusters containing similar words, i.e. words tend to co-occur in the same document. Hence the self-organizing map may measure the word co-occurrence similarity among documents. Examples of F1 measures on SOM-based systems (Category Art) are shown in Table 1 in our SOM-based text categorization experiment. The Micro-F1 and Macro-F1 measures on the SOM-based systems for text categorization are illustrated in Table 2.

Table 1
Examples of F1 measures on SOM-based systems (category art)

Dim & MapSize	F1 (%)
K80-10 × 10B	***94.30***
K80-20 × 20B	93.66
K80-40 × 40B	89.32
K80-Average	92.43%
K100-10 × 10B	88.46
K100-20 × 20B	94.17
K100-40 × 50B	85.19
K100-Average	89.27

Table 2
Micro-F1 and macro-F1 measures on SOM-based systems for text categorization

Dim & MapSize	Micro-F1 (%)	Macro-F1 (%)
K80-10 × 10B	84.68	84.68
K80-20 × 20B	***85.71***	***85.71***
K80-40 × 40B	82.21	82.21
K80-Average	84.20	84.20
K100-10 × 10B	84.14	83.91
K100-20 × 20B	84.28	84.28
K100-40 × 50B	82.51	82.51
K100-Average	83.64	83.57

5.2. Experiments with the LSI-based MTC system

We used 6000 documents in total. Among these, 3000 of them were labeled "Pos" and 3000 were labeled "Neg". We randomly select 80% "Pos" documents in corpora as the training documents and the remainder as testing documents. Both of the "Pos" or "Neg" documents contain topics of "Astronomy", "Art", "Economy", "Medicine", "Physics" and "Politics". The experimental results will be discussed later. The sources of the corpora are shown in Table 3.

Table 3
Sources of the corpora

Source	Website
Grolier online	http://go-passport.grolier.com/
Science American	http://www.sciam.com.tw/
Yahoo!	http://www.yahoo.com
UDN News	http://udn.com/NEWS/main.html
CNA News	http://www.cna.com.tw/
Taiwan Panorama	http://www.taiwan-panorama.com/

Please cite this article in press as: Lee, C. -H. , & Yang, H.-C., Construction of supervised and unsupervised learning systems ..., *Expert Systems with Applications* (2008), doi:10.1016/j.eswa.2007.12.052

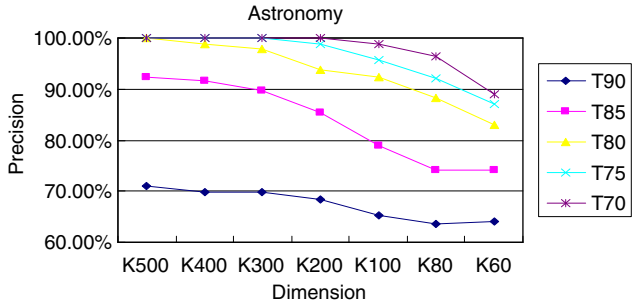

Fig. 7. Precision on class astronomy.

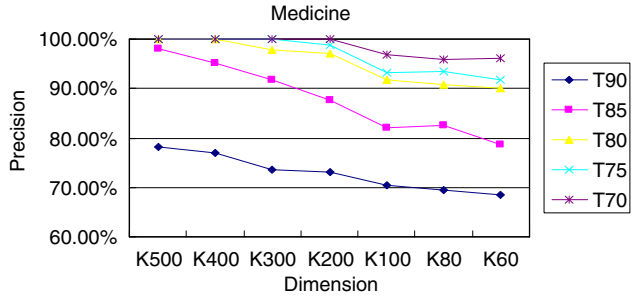

Fig. 8. Precision on class medicine.

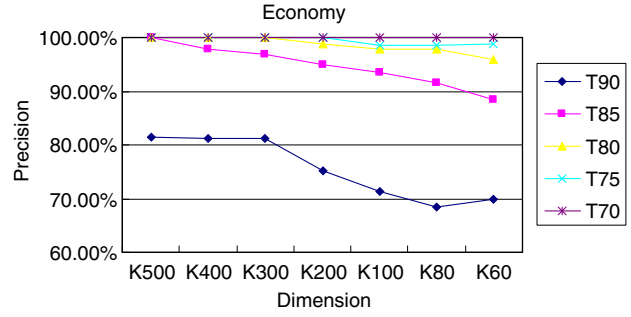

Fig. 9. Precision on class economy.

Experimental results were shown in Figs. 7–9. We obtain higher precision with a lower coverage angle. Figs. 10–12 indicates measurements of F1. In F1 measurement, the data showed that higher coverage angle will lead to more stable results. According to the above results, we

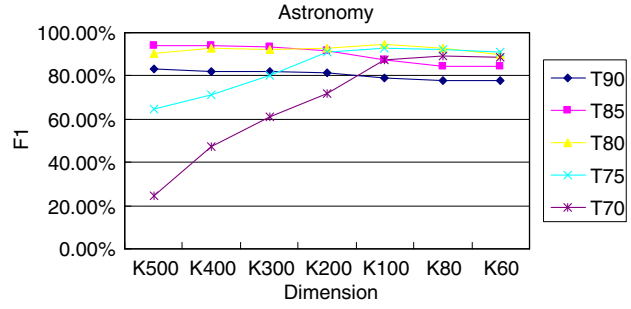

Fig. 10. F1 on class astronomy.

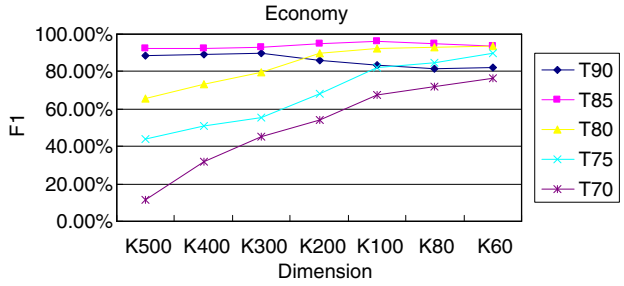

Fig. 11. F1 on class economy.

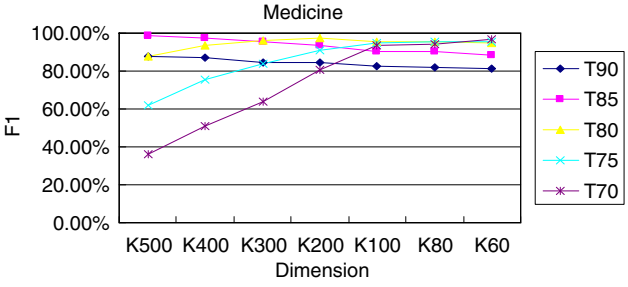

Fig. 12. F1 on class medicine.

can obtain a stable and high precision in multi-language text classification with: (1) coverage angle is setting between 80° and 90°, and (2) The dimension of LSI space is selected below dimensions of 100.

Experimental result indicated that the performance on the precision, recall and F1 are quite good using LSI technique to categorize the multi-language text. The F1 measurement has an average value of 70% and the precision can reach 80% using our algorithm.

5.3. Experiments with the SVM-based MTC system

In this section, the evaluation results of the SVM-based MTC system are discussed. We used labeled documents to train our classifier system, and unlabeled documents to evaluate performance of system. We compare performance of the six classifier results by accuracy, recall, precision and F1 measures. As shown in Figs. 13–16, the measures of the developed SVM classifiers with Linear SVM, Gaussian

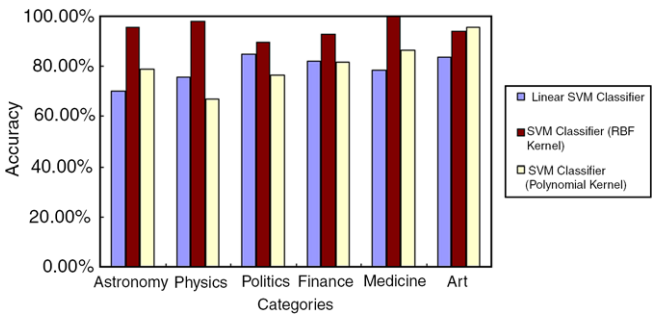

Fig. 13. Results of accuracy rate of developed SVM classifiers.

Please cite this article in press as: Lee, C. -H. , & Yang, H.-C., Construction of supervised and unsupervised learning systems ..., *Expert Systems with Applications* (2008), doi:10.1016/j.eswa.2007.12.052

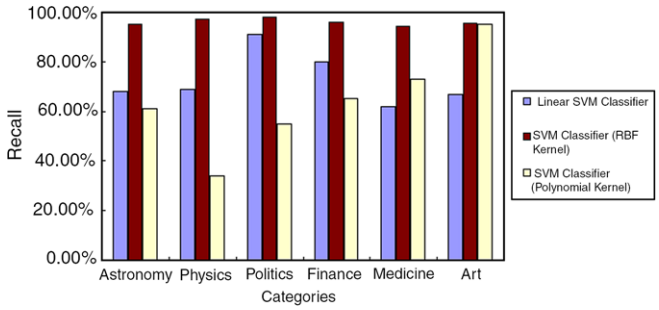

Fig. 14. Results of recall measures of developed SVM classifiers.

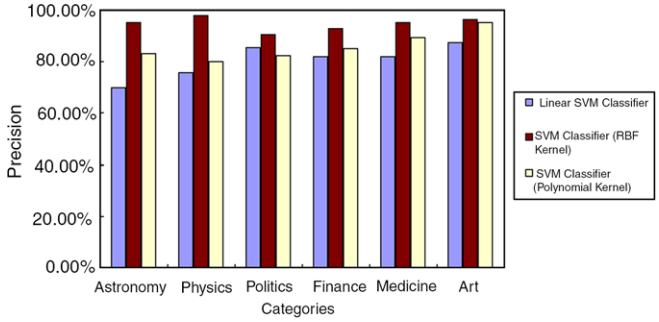

Fig. 15. Results of precision measures of developed SVM classifiers.

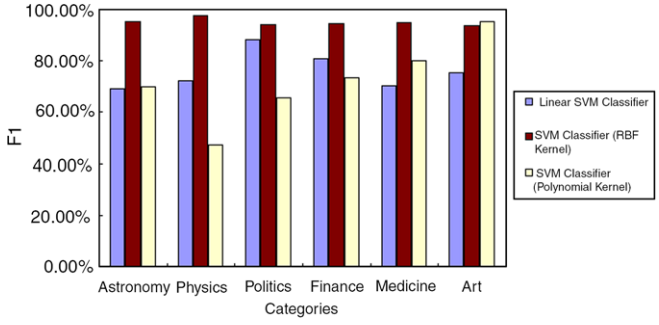

Fig. 16. Results of F1 measures of developed SVM classifiers.

Radial Basis Function (RBF) and Polynomial kernels are illustrated.

6. Performance evaluation with a comparative study

6.1. Comparison of performance

The performance difference in MTC between the LSI-based system and the SOM-based system may not look significant because in essence the above unsupervised techniques utilized for text categorization are basically based on techniques of dimensionality reduction of text feature space, and to some extent the training processes in both unsupervised approaches are similar. As a result, in this work we did not evaluate the performance difference between these two methods. Instead, we select the implemented LSI-based MTC system as a representative of unsupervised methods, and concentrate on comparing the performance between the unsupervised (i.e. the LSI-based method) and the supervised (i.e. the SVM-based method) for fulfilling the MTC tasks. For comparing the performance of these two techniques, in this work we used a unified collected corpus (including 600 Chinese texts and 600

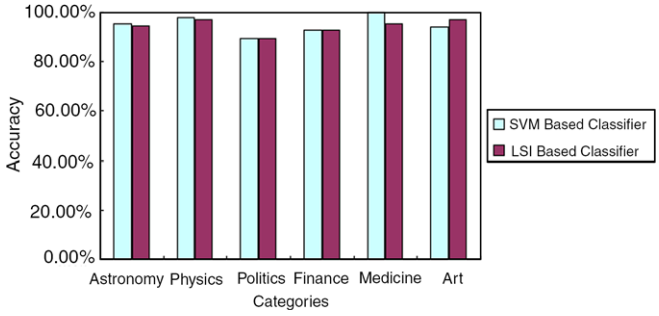

Fig. 17. Results of comparative study: accuracy measures.

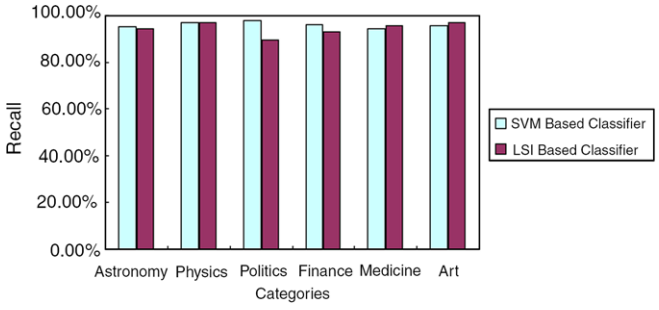

Fig. 18. Results of comparative study: recall measures.

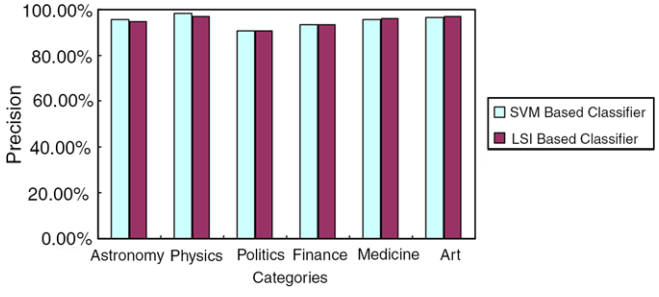

Fig. 19. Results of comparative study: precision measures.

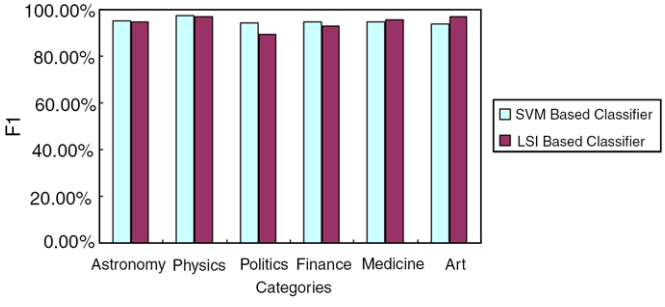

Fig. 20. Results of comparative study: F1 measures.

English texts) to test the classification systems. The experimental results including accuracy, recall, precision and F1 measures are shown in Figs. 17–20.

6.2. Discussion

The experimental results in the previous section suggest that, compared with the SVM-based supervised technique the LSI-based unsupervised technique generally achieve excellent overall performance, although its resulting performance was slightly behind the SVM-based supervised method. However, supervised text categorization requires the extra effort to predefine the categories and to assign category labels to the documents in the training set. This can be very tedious in a huge and dynamic text databases. Also, for a supervised categorization, different human experts may disagree when deciding under which category to categorize a given document. This leads us to believe that by nature the ideal multilingual text categorization should be an unsupervised task rather than a supervised one.

7. Related work

Recent studies in comparing the performance of different categorization techniques have been based largely on experimental approaches (Almuallim & Dietterich, 1994; Dietterich, Hild, & Bakiri, 1995; Wettschereck & Dietterich, 1995). Empirical comparisons among different algorithms suggest that no single method is best for all learning tasks (Salzberg, 1991; Shavlik, Mooney, & Towell, 1991). In other words, each method is best for some, but not all tasks.

In this paper we present the development and performances evaluation of the leading supervised and unsupervised approaches for multilingual text categorization by using various performance measures. A number of text categorization techniques have been developed in recent years. Generally speaking, these techniques can be categorized into two groups: unsupervised learning methods (e.g. LSI approaches) and supervised learning methods (e.g. SVM techniques). Multilingual text categorization is a relatively new research topic, about which not much previous work in the literature appears to be available. Adeva, Calvo, and Ipiña (2005) provided a review of methods related to multilingual (Spanish and Basque) text categorization. They compared different feature extraction strategies such as *n*-gram-based stemming and classic stemming in preprocessing of multilingual documents. On the other hand, they also compared performance of different classification methods in multilingual text categorization such as naïve Bayes, Rocchio and k-nearest neighbor. Jalam, Clech, and Rakotomalala (2004) proposed an original framework for multilingual (English, French, and German) text categorization. Their framework contains two new steps including language identify and language translation. They applied their framework to classifying news information which were written in different languages. For the solutions of supervised techniques, many learning algorithms such as k-nearest neighbor (k-NN) (Masand et al., 1992; Yang & Pedersen, 1997; Yang & Liu, 1999), support vector machines (SVM) (Joachims, 1998), neural networks (Wiener et al., 1995; Ng et al., 1997), linear least squares fit (LLSF) (Yang & Liu, 1999) and naïve Bayes (NB) (Koller et al., 1997; McCallum & Nigam, 1998) have been applied to text classification. A comparison of these techniques is addressed by Yang and Liu (1999). On the other hand, for unsupervised solutions, LSI techniques are well-known approaches for solving information retrieval problems. They were used to tackle the issues of indexed terms containing synonymy and polysemy in the process of text retrieval.

8. Conclusion

Multilingual text categorization is a challenging task. In this work we implemented and measured the performance of the leading supervised and unsupervised approaches for multilingual text categorization. We selected support vector machines (SVM), latent semantic indexing (LSI) and self-organizing maps (SOM) techniques for system implementation. We have shown how our developed platform models including unsupervised and supervised learning systems can provide respectable accuracy in multilingual text categorization. The major conclusion from the joint experimentation is that the methods are possibly complementary. An integrated system developed to overcome the disadvantages of each approach will give better results. We suggest to implement the hybrid system, which will mainly contribute in the following directions: (1) reducing manual effort in the supervised system: since the unsupervised system disregards the categories of the texts during the clustering process, thus human effort is reduced. (2) Also, in future work we want to see if the results can be generalized to other languages, e.g. French, and Japanese. If the results were positive, a generic algorithm would be found that worked well on nearly any language.

References

Adeva, J. J. G., Calvo, R. A., & Ipiña, D. L. D. (2005). Multilingual approaches to text categorisation. *UPGRADE: The European Journal for the Informatics Professional, 6*(3), 43–51.

Almuallim, H., & Dietterich, T. G. (1994). Learning Boolean concepts in the presence of many irrelevant features. *Artificial Intelligence, 69*, 279–305.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391–407.

Dietterich, T. G., Hild, H., & Bakiri, G. (1995). A comparison of ID3 and backpropagation for english text-to-speech mapping. *Machine Learning, 18*, 51–80.

Jalam, R., Clech, J., & Rakotomalala, R. (2004). Cadre pour la catégorisation de textes multilingues. In C. Fairon, G. Prunelle, & A. Dister (Eds.), *7èmes Journées internationales d'Analyse statistique des Données Textuelles* (pp. 650–660). Belgique, Marsh: Louvain-la-Neuve.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. In *Proceedings of the 10th European conference on machine learning* (pp. 137–142). Springer.

Kohonen, T. (1982). Self-organizing formation of topologically correct feature maps. *Biological Cybernetics, 43*, 59–69.

Kohonen, T. (1995). *Self-organizing maps*. Berlin: Springer-Verlag.

Koller, D., & Sahami, M. (1997). Hierarchically classifying documents using very few words. In *Proceedings of 14th international conference on machine learning, Nashville, USA* (pp. 170–178).

Lee., C. H., & Yang, H. C. (2004). A text mining approach for text categorization via computing semantic relatedness using support vector machines. In *Proceedings of the 2004 IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC 2004)*.

Lee, C. H., Yang, H. C., Hsu, F. C., Chen, T. C., & Hung, C. C. (2005). A multiple classifier approach for measuring text relatedness based on support vector machines techniques. In *Proceedings of the 9th world multi-conference on systemics, cybernetics and informatics (WMSCI 2005), Orlando, USA*.

Lee, C. H., & Yang, H. C. (2003). A multilingual text mining approach based on self-organizing maps. *Applied Intelligence, 18*(3), 295–310.

Lee, C. H., & Yang, H. C. (2005). A classifier-based text mining approach for evaluating semantic relatedness using support vector machines. In *International conference on information technology (ITCC 2005)*. Las Vegas, Nevada, USA: IEEE Computer Society.

Littman, M. L., Dumais, S. T., & Landauer, T. K. (1998). Automatic cross-language information retrieval using latent semantic indexing. In G. Grefenstette (Ed.), *Language Information Retrieval* (pp. 51–62). Kluwer Academic Publish.

Masand, B., Linoff, G., & Waltz, D. (1992). Classifying news stories using memory based reasoning. In *Proceedings of the 15th ACM SIGIR conference on research and development in information retrieval* (pp. 59–64).

McCallum, A., & Nigam. K. (1998). A comparison of event models for naive Bayes text classification. In *AAAI-98 workshop on learning for text categorization*.

Ng, H. T., Goh, W. B., & Low, K. L. (1997). Feature selection, perceptron learning, and a usability case study for text categorization. In *Proceedings of the 20th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 67–73).

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill Book Company.

Salzberg, S. (1991). A nearest hyperrectangle learning method. *Machine Learning, 6*, 277–309.

Shavlik, J. W., Mooney, R. J., & Towell, G. G. (1991). Symbolic and neural learning algorithms: An experimental comparison. *Machine Learning, 6*, 111–144.

Vapnik, V. (1995). *The nature of statistical learning theory*. NY: Springer.

Vapnik, V. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Nets, 10*(5), 988–999.

Wettschereck, D., & Dietterich, T. G. (1995). An experimental comparison of the nearest-neighbor and nearest-hyperrectangle algorithms. *Machine Learning, 19*, 5–27.

Wiener, E., Pedersen, J. O., & Weigend, A. S. (1995). A neural network approach to topic spotting. In *Proceedings of the fourth annual symposium on document analysis and information retrieval*.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval, 1*(1–2), 69–90.

Yang, H. C., & Lee, C. H. (2005a). A text mining approach for automatic construction of hypertexts. *Expert Systems with Applications, 29*(4), 723–734.

Yang, H. C., & Lee, C. H. (2005b). Automatic category theme identification and hierarchy generation for Chinese text categorization. *Journal of Intelligent Information Systems, 25*(1), 47–67.

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the 22nd ACM international conference on research and development in information retrieval, Berkeley, USA*.

Yang, Y., & Pedersen, J. P. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the 14th international conference on machine learning* (pp. 412–420).

Yang, H. C., & Lee, C. H. (2004). A text mining approach on automatic generation of web directories and hierarchies. *Expert Systems with Applications, 27*(4), 645–663.

Yang, H. C., & Lee, C. H. (2008). Image semantics discovery from web pages for semantic-based image retrieval using self-organizing maps. *Expert Systems with Applications: An International Journal, 34*(1), 266–279.

Please cite this article in press as: Lee, C. -H. , & Yang, H.-C., Construction of supervised and unsupervised learning systems ..., *Expert Systems with Applications* (2008), doi:10.1016/j.eswa.2007.12.052