Feature Reinforcement Approach to Poly-lingual Text Categorization

Chih-Ping Wei[1], Huihua Shi[2], and Christopher C. Yang[3]

[1] Institute of Technology Management, National Tsing Hua University, Taiwan, ROC
[2] Department of Information Management, National Sun Yat-sen University, Taiwan, ROC
[3] Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong

Abstract. With the rapid emergence and proliferation of Internet and the trend of globalization, a tremendous amount of textual documents written in different languages are electronically accessible online. Poly-lingual text categorization (PLTC) refers to the automatic learning of a text categorization model(s) from a set of preclassified training documents written in different languages and the subsequent assignment of unclassified poly-lingual documents to predefined categories on the basis of the induced text categorization model(s). Although PLTC can be approached as multiple independent monolingual text categorization problems, this naïve approach employs only the training documents of the same language to construct a monolingual classifier and fails to utilize the opportunity offered by poly-lingual training documents. In this study, we propose a feature reinforcement approach to PLTC that takes into account the training documents of all languages when constructing a monolingual classifier for a specific language. Using the independent monolingual text categorization (MnTC) technique as performance benchmarks, our empirical evaluation results show that the proposed PLTC technique achieves higher classification accuracy than the benchmark technique does in both English and Chinese corpora.

1 Introduction

With advances in information and networking technologies, organizations have been actively gathering competitive intelligence information from various online sources, and facilitating information and knowledge sharing within or beyond organizational boundaries. Such e-commerce and knowledge management applications generate and maintain a tremendous amount of textual documents in organizational repositories. To facilitate subsequent access to these documents, use of categories to manage this ever-increasing volume of documents is often observed at both organizational and individual levels. Text categorization deals with the assignment of documents to appropriate categories on the basis of their contents [1][5][6][17]. Central to text categorization is the automatic learning of a text categorization model from a training set of preclassified documents. The induced text categorization model then can classify (or predict) the particular category (or categories) to which a new document belongs.

Various text categorization techniques have been proposed [1][5][6][8][16][17]; however, most of them focus on monolingual documents (i.e., all documents are

D.H.-L. Goh et al. (Eds.): ICADL 2007, LNCS 4822, pp. 99–108, 2007.

written in the same language) in both the learning of a text categorization model and the category assignment of new documents. Because of the trend of globalization, an organization or individual often generates, acquires, and then archives documents written in different languages (i.e., poly-lingual documents). Besides, many countries adopt multiple languages as their official languages. Assume the languages involved in a repository include L_1, L_2, …, L_s, where $s \geq 2$. That is, the set of ploy-lingual documents contains some documents in L_1, some in L_2, …, and some in L_s. If an organization or individual has already organized these poly-lingual documents into existing categories and would like to use this set of precategorized documents as training documents for constructing text categorization models to classify into appropriate categories newly arrived poly-lingual documents, the organization and individual faces the poly-lingual text categorization (PLTC) problem.

PLTC can adopt a naïve approach by considering the problem as multiple independent monolingual text categorization problems. The naïve approach only employs the training documents of a language to construct a monolingual classifier of the same language and ignores all training documents of other languages. When a new document in a specific language arrives, we select the corresponding classifier to predict appropriate category(s) for the target document. However, this independent construction of each monolingual classifier fails to utilize the opportunity offered by poly-lingual training documents to improve the effectiveness of the classifier when the representativeness of the training documents of another language is higher.

For multilingual text categorization, some prior studies address the challenge of cross-lingual text categorization (i.e., learning from a set of training documents written in one language and then classifying new documents in a different language) [3][13]. However, prior research has not paid much attention to PLTC yet. This study is motivated by the importance of providing PLTC support to organizations and individuals in the increasingly globalized and multilingual environment. Specifically, we propose a PLTC technique that takes into account all training documents of all languages when constructing a monolingual classifier for a specific language. For purposes of the intended feasibility assessment and illustration, this study concentrates on only two languages involved in poly-lingual documents and deals with single-category documents rather than multi-category documents. To support linguistic interoperability between training documents in different languages, we rely on a statistical-based bilingual thesaurus that is constructed automatically from a collection of parallel documents. Experimentally, we evaluate the effectiveness of the proposed PLTC technique using independent monolingual classifiers built via the aforementioned naïve approach as performance benchmarks.

2 Literature Review

Text categorization refers to the assignment of documents, on the basis of their contents, to one or more predefined categories. Many text categorization techniques have been proposed [1][5][6][8][16][17], but most of them focus on monolingual documents. Central to text categorization is the automatic learning of a text categorization model from a set of preclassified documents that serve as training examples. The resulting categorization model will then be used to classify the

particular category or categories to which an unclassified document belongs. The process of (monolingual) text categorization generally includes three main phases: feature extraction and selection, document representation, and induction [1][12].

Feature extraction extracts terms (or features) from the training documents. However, different languages exhibit different grammatical and lexical characteristics that significantly affect how the features in documents are segmented. Feature selection reduces the size of the feature space. Popular feature selection techniques include TF (term frequency), TF×IDF (IDF denotes inverse document frequency), correlation coefficient, χ^2 metric, and mutual information [6][7][10].

In the document representation phase, each document is represented by a vector space jointly defined by the top-*k* features selected in the previous phase and, in the meanwhile, is labeled to indicate its category membership. Binary (which indicates the presence or absence of a feature in a document), within-document TF, and TF×IDF are the most popular representation methods.

In the induction phase, a text categorization model(s) that distinguishes categories from one another, on the basis of the set of training documents is constructed. Prevalent learning techniques employed for text categorization include decision-tree induction, decision-rule induction [1][5], k-nearest neighbor (kNN) classification [8][16], neural network, Naïve Bayes probabilistic classifier [2][8], SVM [6], and statistical approach [17]. Sebastiani [9] offer empirical evaluations of different learning techniques for text categorization.

Cross-lingual text categorization (CLTC) deals with learning from a set of training documents (i.e., the training corpus) written in one language and then classifying unclassified documents (i.e., the prediction corpus) in a different language [3][4][13]. The major challenge facing CLTC is cross-lingual semantic interoperability that establishes the bridge between the representations of the training and prediction documents that are written in different languages. Although several studies have been conducted on CLTC, CLTC does not take poly-lingual preclassified documents as training examples as PLTC does. Therefore, CLTC is not able to take the advantage of the semantics embedded in poly-lingual training documents for text categorization model learning but relies on monolingual training documents and a translation mechanism to classify new documents in another language. In this work, we focus on PLTC with the support of automatic multilingual thesaurus.

3 Poly-lingual Text Categorization (PLTC) with Feature Reinforcement

In this study, we propose a feature reinforcement approach to PLTC with the support of an automatic multilingual (or bilingual when $s = 2$) thesaurus to address potential limitations of the naïve approach. Figure 1 shows the overall design of the proposed PLTC technique, which consists of three main tasks: 1) bilingual thesaurus construction for building a statistical bilingual thesaurus (in this study, English and Chinese) from a parallel corpus, 2) categorization learning for inducing a text categorization model for each language based on a set of preclassified documents in languages L_1 and L_2, and 3) category assignment for predicting appropriate categories for unclassified documents in either L_1 or L_2.

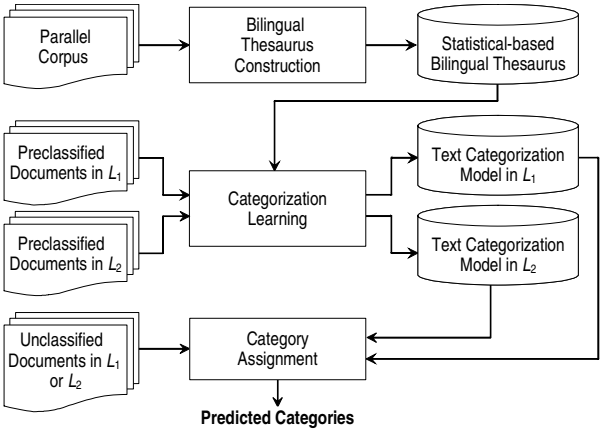

Fig. 1. Overall Process of Our Proposed PLTC Technique

3.1 Bilingual Thesaurus Construction

This task automatically constructs a statistical-based bilingual thesaurus using the co-occurrence analysis technique [15], commonly employed in cross-lingual information retrieval (CLIR) and CLTC research. Given a parallel corpus, the thesaurus construction process starts from term extraction and selection. In this study, we deal with only English and Chinese documents. We use the rule-based part-of-speech tagger [4] to tag each word in the English documents in the parallel corpus and then adopt the approach proposed by Voutilainen [11] to extract noun phrases from the syntactically tagged English documents. For the Chinese documents in the parallel corpus, we employ a hybrid of dictionary-based and statistical approaches to extract Chinese terms [14][15].

Subsequently, we adopt the TF×IDF scheme proposed by Yang and Luk [14] as the term selection metric. The term weight of a term f_j (English or Chinese) in a parallel document d_i (denoted as tw_{ij}) is calculated as:

$$tw_{ij} = tf_{ij} \times \log\left(\frac{N_P}{n_j} \times l_j\right)$$

where tf_{ij} is the term frequency of f_j in d_i, N_P is the total number of parallel documents in the corpus, n_j is the number of parallel documents in which f_j appears, and l_j is the length of f_j (where l_j denotes the number of English words if f_j is an English term or the number of Chinese characters if f_j is a Chinese term).

For each parallel document, the top k_{clt} English and k_{clt} Chinese terms with the highest TF×IDF values (i.e., tw_{ij}) and that simultaneously occur in more than δ_{DF} documents are selected for each parallel document. On the basis of the concept that relevant terms often co-occur in the same parallel documents, the co-occurrence analysis first measures the co-importance weight cw_{ijh} between terms f_j and f_h in the parallel document d_i as follows [14]:

$$cw_{ijh} = tf_{ijh} \times \log\left(\frac{N_P}{n_{jh}}\right)$$

where tf_{ijh} is the minimum of tf_{ij} and tf_{ih} in d_i, and n_{jh} is the number of parallel documents in which both f_j and f_h occur.

Finally, the relevance weights between f_j and f_h are computed asymmetrically as follows [14]:

$$rw_{jh} = \frac{\sum_{i=1}^{N_P} cw_{ijh}}{\sum_{i=1}^{N_P} tw_{ij}} \text{ and } rw_{hj} = \frac{\sum_{i=1}^{N_P} cw_{ijh}}{\sum_{i=1}^{N_P} tw_{ih}}$$

where rw_{jh} (or rw_{hj}) denotes the relevance weight from f_j to f_h (or from f_h to f_j).

After we estimate all directional statistical strengths between each pair of English and Chinese terms selected by the term extraction and selection phase, pruning of insignificant strengths is performed. Specifically, if the statistical strength from one term to another is less than a relevance threshold δ_{rw}, we remove the link. As a result, we construct a statistical-based bilingual thesaurus from the input parallel corpus.

3.2 Categorization Learning

The categorization learning task is an important component of our proposed PLTC technique. As we show in Figure 2, when training a monolingual classifier for language L_i (L_1 or L_2), our proposed categorization learning method takes into account not only the preclassified documents in L_i but also the preclassified documents in another language L_j as well as the statistical-based bilingual thesaurus. Specifically, to train a monolingual classifier for L_i, the categorization learning task involves four phases: feature extraction (for L_i and L_j), feature reinforcement and selection (for L_i), document representation (in L_i), and induction.

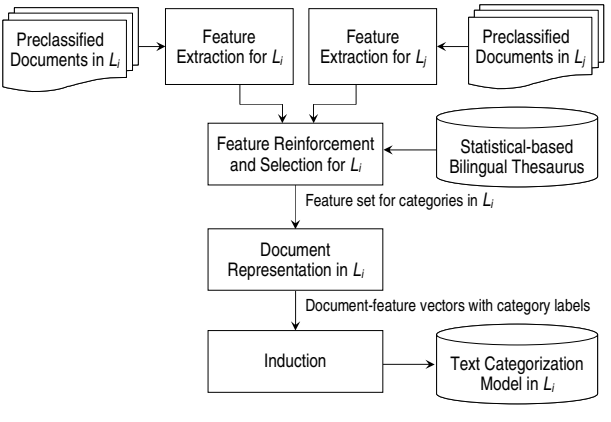

Fig. 2. Process of Categorization Learning (for L_i)

Feature Extraction: In this phase, we extract features from the preclassified documents in both languages. We employ the same feature extraction mechanisms as those in the bilingual thesaurus construction task to extract as features nouns and noun phrases

from preclassified English documents and Chinese terms from preclassified Chinese documents.

Feature Reinforcement and Selection: We assess the discriminating power of each feature in its respective training corpus and language. In this work, we adopt the χ^2 statistic metric, which measures the dependence between a feature f and a category C_i (denoted as $\chi^2(f, C_i)$). Using a two-way contingency table of f and C_i, let n_{r+} be the number of documents in C_i in which f occurs, n_{r-} be the number of documents in C_i in which f does not appear, n_{n+} be the number of documents in categories other than C_i in which f occurs, n_{n-} be the number of documents in categories other than C_i in which f does not appear, and n be the total number of documents under discussion. The χ^2 statistic of f relevant to C_i thus is defined as [18]:

$$\chi^2(f, C_i) = \frac{n \times (n_{r+}\, n_{n-} - n_{r-}\, n_{n+})^2}{(n_{r+} + n_{r-})(n_{n+} + n_{n-})(n_{r+} + n_{n+})(n_{r-} + n_{n-})}$$

Once the χ^2 statistic of the feature f relevant to each category C_i is derived, the overall χ^2 statistic of f for all categories T is calculated using the weighted average scheme [18]. That is,

$$\chi^2(f, T) = \sum_{C_i \in T} p(C_i) \times \chi^2(f, C_i)$$

where $p(C_i)$ is the number of documents in C_i divided by n.

After the χ^2 statistic scores for all features in both languages are obtained, we start to reassess the discriminating power of a feature in one language by considering the discriminating power of its related features in another language. The reason for such crosschecking between two languages is that if a feature in one language and its related features in another language are having high χ^2 statistic scores, it is likely that the feature has greater discriminatory power. However, inconsistent assessments between two languages (i.e., the χ^2 statistic score of a feature is high in one language but the χ^2 statistic scores of its related features are low in another language) will result in a lower confidence on the discriminatory power of the feature. In this work, we adopt this crosschecking process as feature reinforcement.

Assume a total of N_1 features in L_1 are extracted from the preclassified training documents (in L_1) and N_2 features in L_2 are extracted from the preclassified training documents (in L_2). Given a feature f_i in L_1, let $R(f_i)$ be the set of features in L_2 that have direct cross-lingual associations to f_i according to the statistical-based bilingual thesaurus derived previously. The alignment weight for f_i in L_1 (denoted as $aw(f_i)$) from its related features (i.e., $R(f_i)$) in L_2 is defined as follows:

$$aw(f_i) = \frac{\sum_{\forall g_j \in R(f_i)} \chi^2(g_j) \times rw_{g_j f_i}}{|R(f_i)|} \times \log \frac{N_2}{|R(f_i)|}$$

where $\chi^2(g_j)$ is the χ^2 statistic score of feature g_j, and $rw_{g_j f_i}$ is the relevance weight from g_j to f_i as specified in the statistical-based bilingual thesaurus.

Subsequently, we use the following formula to arrive at the overall weight of a feature f_i by combining the weights of f_i derived from the training documents in both languages:

$$w(f_i) = \alpha \times \chi^2(f_i) + (1-\alpha) \times aw(f_i)$$

where α denotes the tradeoff between the χ^2 statistic score of f_i in its original language and the alignment weight of f_i derived from the other language (where $0 \leq \alpha \leq 1$).

Once the overall weights of all features are derived for both languages, we then perform feature selection. For each language (L_1 or L_2), we select the k features with the highest overall weights as features to represent each training document of the respective language.

Document Representation: In this phase, the training documents of each language are represented using the corresponding feature set selected previously. In this study, we consider three prevalent document representation schemes that include binary, within-document TF and TF×IDF and empirically evaluate their effects on classification effectiveness. That is, each training document d_i forms a document-feature vector $\vec{d}_i$.

Induction: The induction phase is to induce two monolingual text categorization models from the preclassified documents in L_1 and L_2, respectively. We adopt the Naïve Bayes probabilistic classifier and Support Vector Machine (SVM) as alternative learning algorithms because of their popularity in prior research on text categorization. The Naïve Bayes classifier uses the joint probabilities of words and categories to estimate the probabilities of categories fitting a particular document. In contrast, SVM is based on Structural Risk Minimization principle and defined over a vector space where the classification or categorization problem is to find a decision surface that best separates the positive and negative training examples with the maximum margin.

3.3 Category Assignment

In the category assignment task, we categorize each unclassified document in L_1 or L_2 using the corresponding text categorization model induced previously. According to the language used in the unclassified document, we use the respective feature extraction method (described in Section 3.1) to extract features from the unclassified document and employ binary, within-document TF, or TF×IDF representation scheme to represent the target unclassified document. Finally, the feature vector of the document is used to determine an appropriate category on the basis of the corresponding text categorization model.

4 Empirical Evaluation

In this section, we report our empirical evaluation of the proposed PLTC approach. In the following subsections, we detail the design of our empirical experiments, including data collection, evaluation procedure and criteria, and our benchmark technique. Subsequently, we discuss important evaluation results.

4.1 Data Collection

To construct a statistical-based bilingual thesaurus requires the parallel documents in two languages. News presses from Government Information Center, Hong Kong

Special Administrative Region of The People's Republic of China (accessible at http://www.info.gov.hk/) were collected for constructing a statistical-based bilingual thesaurus. Specifically, the parallel corpus collected for our experimental purpose contains 2074 pairs of Chinese and English news presses.

Two additional monolingual document corpora were collected for evaluating the effectiveness of our proposed PLTC technique. The English and Chinese corpora are news presses collected from Government Information Center, Hong Kong. Both the English and Chinese corpora consist of 278 news presses related to eight categories. We merge these two monolingual corpora into a poly-lingual corpus for our evaluation purpose.

4.2 Evaluation Procedure and Criteria

To evaluate the effectiveness of PLTC, we randomly select 50% of the documents in the English and the Chinese corpora respectively as our training dataset and the remainder 50% of the documents in these two corpora as the testing dataset. To avoid the bias caused by random sampling, we repeat the sampling and train-and-test process 10 times and evaluate the effectiveness of the PLTC technique under investigation by averaging the performance obtained from these 10 individual trials. We measure the effectiveness of PLTC on the basis of classification accuracy, which is defined as the percentage of documents in the testing dataset that the PLTC technique under investigation correctly classifies into the predefined categories.

4.3 Performance Benchmark

As mentioned previously, the PLTC problem can be simply approached as multiple independent monolingual text categorization problems. That is, we construct for each language a monolingual text categorization model (i.e., classifier) on the basis of the training examples of the respective language only. For an unclassified document, we select the corresponding classifier to predict the appropriate category for the target document. In this study, we adopt this naïve approach as our benchmark technique and refer it as the MnTC technique.

4.4 Comparative Evaluation

We first conduct a series of tuning experiments to determine appropriate values for the parameters involved in bilingual thesaurus construction. Our experimental results suggest that setting δ_{DF} as 3, k_{clt} as 30, and δ_{rw} as 0.15 would be appropriate. Thus, these values are adopted for our subsequent experiments. Moreover, we also perform tuning experiments to determine the value for α (required by the PLTC technique). Our results show the best classification accuracy is achieved when α equals to 0.1. Thus, we employ this value for the subsequent comparative evaluation.

As we summarize in Tables 1 and 2, across all representation schemes and learning algorithms examined, our proposed PLTC outperforms the benchmark technique (i.e., MnTC) in both document corpora (i.e., English and Chinese). In addition, the PLTC technique using the Naïve Bayes classifier and binary representation achieves the best classification accuracy (i.e., 72.42% and 71.49%) across two different corpora.

Table 1. Comparison of Effectiveness of PLTC and MnTC on English Corpus

	Representation	Classification Accuracy		Δ
		MnTC	PLTC	
Naïve Bayes	Binary	67.63%	72.42%	4.79%
	TF	68.25%	70.26%	2.01%
	TF×IDF	67.24%	68.54%	1.30%
SVM	Binary	66.14%	68.87%	2.73%
	TF	62.45%	68.97%	6.52%
	TF×IDF	62.59%	68.54%	5.95%

Note: Δ denotes the improvement, calculated as (Classification Accuracy of PLTC – Classification Accuracy of MnTC), in Tables 1–2.

Table 2. Comparison of Effectiveness of PLTC and MnTC on Chinese Corpus

	Representation	Classification Accuracy		Δ
		MnTC	PLTC	
Naïve Bayes	Binary	65.61%	71.49%	5.88%
	TF	64.29%	67.63%	3.34%
	TF×IDF	63.48%	67.34%	3.86%
SVM	Binary	62.33%	65.64%	3.31%
	TF	58.92%	64.12%	5.20%
	TF×IDF	58.68%	63.57%	4.89%

5 Conclusion and Future Research Directions

In this work, we have investigated poly-lingual text categorization (PLTC). Many text categorization techniques have been proposed in the literature; however, most of them deal with monolingual documents. In response, we propose a feature-reinforcement-based PLTC technique that takes into account all training documents of all languages when constructing a monolingual classifier for a specific language. Using the independent monolingual text categorization (MnTC) technique as performance benchmarks, our empirical evaluation results show that our proposed PLTC technique achieves higher classification accuracy than the benchmark technique does in both English and Chinese corpora.

Some future research works related to this study include the following: Our proposed PLTC technique focuses only on two languages. It would be interesting to extend our proposed PLTC technique when the preclassified poly-lingual documents are written in more than two languages. In addition to PLTC, other poly-lingual document management issues (e.g., poly-lingual event detection) require further research attention.

Acknowledgments

This work was supported by the National Science Council of the Republic of China under the grants NSC 93-2416-H-110-021 and NSC 94-2416-H-110-002.

References

[1] Apte, C., Damerau, F., Weiss, S.: Automated Learning of Decision Rules for Text Categorization. ACM Transactions of Information Systems 12(3), 233–251 (1994)

[2] Baker, L.D., Mccallum, A.K.: Distributional Clustering of Words for Text Classification. In: Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998), pp. 96–103 (1998)

[3] Bel, N., Koster, C.H.A., Villegas, M.: Cross-Lingual Text Categorization. In: Koch, T., Sølvberg, I.T. (eds.) ECDL 2003. LNCS, vol. 2769, pp. 126–139. Springer, Heidelberg (2003)

[4] Brill, E.: Some Advances in Rule-Based Part of Speech Tagging. In: Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94), Seattle, WA, pp. 722–727 (1994)

[5] Cohen, W.W., Singer, Y.: Context-sensitive Learning Methods for Text Categorization. ACM Transactions on Information Systems, 17(2), 141–173 (1999)

[6] Dumais, S., Platt, J., Heckerman, D., Sahami, M.: Inductive Learning Algorithms and Representation for Text Categorization. In: Proceedings of the 1998 ACM 7th International Conference on Information and Knowledge Management (CIKM 1998), pp. 148–155 (1998)

[7] Lam, W., Ho, C.Y.: Using A Generalized Instance Set for Automatic Text Categorization. Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998), 81–89 (1998)

[8] Larkey, L., Croft, W.: Combining Classifiers in Text Categorization. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1996), Zurich, Switzerland, pp. 289–297 (August 1996)

[9] Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34(1), 1–47 (2002)

[10] Schutze, H., Hull, D.A., Pedersen, J.O.: A Comparison of Classifiers and Document Representations for the Routing Problem. In: Proceedings of the 18th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 229–237 (1995)

[11] Voutilainen, A.: Nptool: A Detector of English Noun Phrases. In: Proceedings of Workshop on Very Large Corpora, Ohio, pp. 48–57 (June 1993)

[12] Wei, C., Hu, P., Dong, Y.X.: Managing Document Categories in E-Commerce Environments: An Evolution-Based Approach. European Journal of Information Systems 11(3), 208–222 (2002)

[13] Wei, C., Lin, Y. T., Yang, C. C.: Cross-Lingual Text Categorization for Global Knowledge Management, Working Paper, Institute of Technology Management, National Tsing Hua University, Hsinchu, Taiwan, R.O.C. (June 2005)

[14] Yang, C.C., Luk, J.: Automatic Generation of English/Chinese Thesaurus Based on a Parallel Corpus in Laws. Journal of the American Society for Information Science and Technology 54(7), 671–682 (2003)

[15] Yang, C.C., Luk, J., Yung, S., Yen, J.: Combination and Boundary Detection Approach for Chinese Indexing. Journal of the American Society for Information Science 51(4), 340–351 (2000)

[16] Yang, Y.: Expert Network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval. In: Proceedings of the 17th International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, pp. 13–22 (July 1994)

[17] Yang, Y., Chute, C.G.: An Example-Based Mapping Method for Text Categorization and Retrieval. ACM Transaction on Information Systems 12(3), 252–277 (1994)

[18] Yang, Y., Pedersen, J.O.: A Comparative Study on Feature Selection in Text Categorization. In: Proceedings of 14th International Conference on Machine Learning, pp. 412–420 (1997)