Research on Text Categorization Based on a Weakly-Supervised Transfer Learning Method

Dequan Zheng, Chenghe Zhang, Geli Fei, and Tiejun Zhao

MOE-MS Key Laboratory of Natural Language Processing and Speech
Harbin Institute of Technology, Harbin, China
dqzheng2007@gmail.com

Abstract. This paper presents a weakly-supervised transfer learning based text categorization method, which does not need to tag new training documents when facing classification tasks in new area. Instead, we can take use of the already tagged documents in other domains to accomplish the automatic categorization task. By extracting linguistic information such as part-of-speech, semantic, co-occurrence of keywords, we construct a domain-adaptive transfer knowledge base. Relation experiments show that, the presented method improved the performance of text categorization on traditional corpus, and our results were only about 5% lower than the baseline on cross-domain classification tasks. And thus we demonstrate the effectiveness of our method.

Keywords: Transfer learning, Text Categorization.

1 Introduction

With the explosion of the text documents on the web, text classification technique has been playing a more and more essential role in helping people find, collect and organize these data. As a result, the study on how to use computer to classify texts automatically has become a key realm in both natural language processing and artificial intelligence field.

Traditional text classification techniques are usually based on machine learning, which means people have to train a categorization model in advance. However, traditional machine learning methods rely on strong assumptions: the first assumption is that training and testing data set should be evenly distributed; and the other is that they should be in homogeneous feature space. Unfortunately, this is not always true in reality, which may lead to the failure of the text classifier in many cases, such as outdated training data set.

At this time, people have to label a large quantity of texts in new domains to meet the needs of training, but tagging new texts and training new models are extremely expensive and time consuming. From another point of view, if we already have a large volume of labeled texts in one domain, it is wasteful to totally abandon them. So how to make full use of these data is what a method called transfer learning aims to solving. Transfer learning means people may extract transfer knowledge from the data

A. Gelbukh (Ed.): CICLing 2012, Part II, LNCS 7182, pp. 144–156, 2012.

available at present and use them in the future, or extract transfer knowledge from one domain and use it in other domains.

As a result, this paper proposes a novel domain-adaptive transfer learning method, which combines linguistic information and statistics. Through learning from available data or knowledge base at hand, we construct a new transfer knowledge base in heterogeneous feature space without tagging new corpuses.

The rest of this paper is organized as follows. In section 2, we give an introduction of related works. In section 3, we describe our method of acquiring transfer knowledge. In section 4, we describe how our text classifier is implemented and how transfer knowledge is used. In section 5, we introduce the results of our experiments and evaluation method. In section 6, we introduce some of our future work.

2 Related Works

2.1 Transfer Learning

Current research work in the field of transfer learning can be mainly divided into three parts. One is example-based transfer learning in homogenous feature space, one is feature-based transfer learning in homogenous space, another is transfer learning in heterogeneous feature space.

The main idea of example-based transfer learning is that, although the training data set is, to some extent, different from the testing set, there must exists some part of it that is suitable for training a reliable categorization model.

As reported in Pan and Yang's survey [1], many transfer learning methods have already proposed. For example, feature-representation-transfer approaches. In the field of feature-based transfer learning, scholars have proposed several algorithms for learning, such as CoCC [2], TPLSA [3], Spectral Domain-Transfer Learning [4], and self-taught clustering [5]. Its basic idea is that by clustering the target and training data simultaneously to allow the feature representation from the training data to influence the target data through a common set of features. Under the new data representation, classification on the target data can be improved.

Some scholars have proposed a method called translated learning [6,7] to solve the problem of using labeled data from one feature space to enhance the classification of other entirely different learning spaces. An important aspect of translated learning is to build a “bridge” to link one feature space (known as the “source space”) to another space (known as the “target space”) through a translator in order to transfer the knowledge from source to target. The translated learning solution uses a language model to link the class labels to the features in the source spaces, which in turn is translated to the features in the target spaces. Finally, this chain of linkages is completed by tracing back to the instances in the target spaces. Another proposed method is domain adaptation with structural correspondence learning, which plays a 'pivot' features in transfer learning [8].

2.2 Text Categorization

Text categorization aims at automatically classifying text documents into certain predefined categories and is an important technique in processing and organizing significant number of texts. It classifies a text into a most possible category by extracting its features and comparing them with those of the predefined categories and, in consequence, enhances the relationship among different texts.

The invention of text classifier can be traced back to the work of Maron in 1961. In 1970, Salton [9] proposed the VSM (Vector Space Model), which has become a classic model for text categorization. In 1990's, with the rapid development of the Internet and vast volume of texts on it, text classifier accordingly developed at a faster speed. A variety of text categorization methods appeared and the one which is based on machine learning has become a dominant one and achieves a very good result.

Most methods of text categorization come from pattern classification and can be divided into three categories. One is statistics-based, such as Naïve Bayes [10], KNN [11], Centroid-Based Model [12], Regression Model [13], SVM [14], and Maximum Entropy Model [15]; one is based on Neural Network Approach [16]; another is rules-based, such as decision-tree [17] and association rules [18]. And the method that is based on statistics is mainly studied and used nowadays.

3 Transfer Knowledge Acquisition

In order to challenge the conventional assumption of machine learning, and to take full advantage of the available data, we propose a novel transfer learning method, hoping to mine knowledge from available data, and applying this transfer knowledge to heterogeneous feature space, so as to help new tasks in new domains.

We learn keywords' syntactic and semantic use by quantifying their contextual information, such as part-of-speech, semantics, possibility of co-occurrence, and position to form transfer knowledge, so as to help cross-domain machine learning tasks.

This paper proposes a novel strategy to automatically acquire transfer knowledge from existing training data. Since automatic text classification is a typical task that involves machine learning, we use text classification tasks to testify the efficacy of our proposed method.

3.1 Knowledge Acquisition in Homogeneous Feature Spaces

Algorithm 1

Step1: corpus pre-processing

For any Chinese document D we do Chinese word segmentation, POS tagging, and then, we wipe off the word that can do little contribution to the linguistic knowledge, such as preposition, conjunction, auxiliary word and etc.

Step2: establish a temporary text for later processing.

Extract *k* keywords for the text by using TF-IDF method (k<=50). And then, we extract all the sentences that contain these 50 keywords and form a temporary document *D'* for acquiring co-occurrence knowledge.

Step3: Calculate the co-occurrence distance.

For a single keyword in the document *D'*, we consider the keyword as the center, and respectively get the left-side and the right-side co-occurrence distance from keyword to its co-occurrence, which is calculated as (1) and (2) (m, n<=5).

$$C_{li} = \left(\frac{1}{2}\right)^{i-1} B_l \quad \text{(i=1,2, ...,m)} \tag{1}$$

$$C_{rj} = \left(\frac{1}{2}\right)^{j-1} B_r \quad \text{(j=1,2,......,n)} \tag{2}$$

In (1) and (2), B_l and B_r are the importance decay factor from keywords, which is calculated as (3). We consider the closer a word is to a keyword, the more important it is.

$$B_l = \frac{1}{\sum_{i=1}^{m}\left(\frac{1}{2}\right)^{i-1}} \qquad B_r = \frac{1}{\sum_{j=1}^{n}\left(\frac{1}{2}\right)^{j-1}} \tag{3}$$

In (3), m and n represent the number of left and right number of words from a keyword.

Step4: accumulate co-occurrence distance.

For the *ith* co-occurrence of a keyword, we extract its part-of-speech POS, and position L. Then we regard the keyword and its *ith* co-occurrence (co-occurrence$_i$, POS$_i$, L) as a relation pair, and C_i as its co-occurrence distance, which is calculated from L. To a single keyword, we accumulate all the C_i of the relation pairs which have the same co-occurrence$_i$ and POSi and, at the same time, record the accumulation times.

Step5: Calculate the average co-occurrence distance.

We calculate the average value of C_i that appears in corpus known as the average co-occurrence distance $\overline{C_i}$ between the keyword and its co-occurrence (co-occurrence$_i$, POS$_i$, L).

Step6: Build up transfer knowledge base.

When all of documents are learned, all keywords and their co-occurrence information (co-occurrence, POS, $\overline{C}$) compose our transfer knowledge base.

Step7: Build index.

In order to improve the processing speed, for the acquired transfer knowledge base, we use such keys as (keyword, co-occurrence, POS) to build an index for our transfer knowledge base.

3.2 Knowledge Acquisition in Heterogeneous Feature Spaces

Acquiring transfer knowledge from heterogeneous feature spaces means to learn transfer knowledge from available data or knowledge base, and to apply it to different domains.

Traditional machine learning requires that training and testing data should be evenly distributed and in homogeneous feature space. If the feature space of testing data changes dramatically, or changes in domain, the training data fails. So we propose a method for acquiring transfer knowledge in heterogeneous feature space.

Our method is based on people's assumption of natural language: In a certain period of time, words in natural language basically remain their syntactic and semantic use, no matter in what circumstance and context. So in this paper, we assume that before processing documents, we can acquire some keywords that can represent their categories from domain experts. For these keywords, we then acquire transfer knowledge for their respective categories from the existing knowledge base.

Algorithm 2

Step1, to simulate the process of acquiring keywords from domain experts, we extract 50 keywords from each category by using TF-IDF methods.

Step2, check if these keywords exist in existing knowledge bases, extract the part-of-speech and co-occurrence distance of their co-occurrences, and form such relation pair (keyword, POS_i, L). For each keyword, we calculate the times of the same POS, and, at the same time, accumulate co-occurrence distance.

Step3, we divide the accumulated co-occurrence distance by its times and regard this value as the average co-occurrence distance $\overline{C}$. For every keyword and all its co-occurrence's part-of-speech, we consider the relation pair (keyword, POS, $\overline{C}$) as the transfer knowledge of this keyword; all the above mentioned relation pairs of in a category compose the transfer knowledge bank for this category.

Step4, In order to improve the processing speed, for the acquired transfer knowledge bank, we use such keys as (keyword, POS) to build an index for our transfer knowledge bank.

4 Application of Transfer Knowledge

In order to test the feasibility and effectiveness of our transfer learning method, we apply our method to automatic text categorization, which is a typical use in machine learning. We firstly extract transfer knowledge from each category and form a transfer knowledge base according to Algorithm 1 and Algorithm 2. And then we use these transfer knowledge bases as the text classifier. When we have a document to be classified, we will calculate its possibility of belonging to one category based on each category's transfer knowledge base, and choose the category with the highest possibility as the document's final category. Algorithm 3 describes the process of text classification tasks in homogeneous feature space, and Algorithm 4 describes the process of text classification tasks in heterogeneous feature space.

4.1 Text Classification in Homogeneous Feature Space

Algorithm 3

Step1: Documents Pre-processing.

For any Chinese document D, we do Chinese word segmentation, POS tagging and then, wipe off the word that can do little contribution to the linguistic ontology knowledge, such as preposition, conjunction, auxiliary word and etc. Next, we extract k keywords from this document by using the TF-IDF method, and regard these k keywords as the features of this document;

Step2: Get the average co-occurrence distance from transfer knowledge base.

For the *ith* keyword Keyword_i in document D and its relation pair <Keyword_i, (CoexistWord, POS)>, we search them in the *jth* transfer knowledge base we constructed according to Algorithm 1 and add up the average co-occurrence distance, and the result is known as C_j^i;

Step3: Calculate the possibility for a document in the *jth* category.

Repeat step 2 until *k* keywords in document D are calculated. The possibility of belonging to the *jth* category is calculated as (4), known as *Eval_D$_j$*..

$$Eval\_D_j = \sum_{i=1}^{k} C_j^i \tag{4}$$

Step4: Choose a category for documents

After we have calculated the document's possibility of belonging to each category, we put the document into the category with the highest possibility as (5).

$$Eval\_D = Max\left(Eval\_D_j\right) \tag{5}$$

In which *j* represents the total number of categories of training document.

4.2 Text Classification in Heterogeneous Feature Space

Algorithm 4

Step1: Documents Pre-processing.

This process is similar to the Step1 in Algorithm 3;

Step2: Get the average co-occurrence distance from transfer knowledge base.

For the *ith* keyword Keyword_i in document D and its relation pair <Keyword_i, POS>, we search them in the *jth* transfer knowledge bank we constructed according to Algorithm 2 and add up the average co-occurrence distance, and the result is marked as C_j^i;

Step3: Get the possibility value for a document in the *jth* category.

This process is similar to the Step3 in Algorithm 3;

Step4: Decide the final category for a document.

This process is similar to the Step4 in Algorithm 3;

5 Experiments and Analysis

5.1 Experimental Data and Evaluation Methods

We choose three Chinese data sets to evaluate our method of transfer learning. One is 863 corpus (2003), one is 863 corpus (2004), another is Tan corpus [19, 20]. In this paper, we design four groups of experiments to test our method.

In the first group of experiments, we aim to test the effectiveness of our transfer learning method when it is used in homogeneous feature space. We use the 863 corpus (2003) as the training data set, acquiring transfer knowledge and constructing transfer knowledge base according to Algorithm 1, and then using this knowledge base to help classify text documents in the 863 corpus (2004). Since the 863 corpus (2003) and the 863 corpus (2004) consist of the same categories of documents, and they are evenly distributed, so we can consider these two date sets in homogeneous feature space.

In the second group of experiments, although we still test the effectiveness of our transfer learning method when it is used in homogeneous feature space, the documents in training and testing data set are unevenly distributed and come from different sources. In the experiment, we choose the same five categories from both 863 corpus (2003) and Tan corpus, and use the former date set as the training data, and the latter as the testing data set. Although unevenly distributed, these documents belong to the same categories, so we still consider them under homogeneous feature space.

In the third experiment, we aim at testing the effectiveness of our transfer learning method itself. We choose our training and testing data set from the same corpus, and do three-cross validation. We test our method in Tan corpus and the 863 corpus respectively.

In the last experiment, we try to test the effectiveness of transfer learning method among heterogeneous feature spaces. We focus on how to take full advantage of the existing knowledge base when testing data is greatly changed in category or there is no training data available. By acquiring transfer knowledge from the 863 corpus according to Algorithm 2, and applying it to 20 categories in Tan corpus which do not in the 863 corpus, we test our transfer learning method in heterogeneous feature space. In this experiment, the 20 selected categories from the 863 corpus which are used as training data set consist of politics and laws, philosophy, economy, literature, art, biology, architecture, transportation and another 12 categories. However, the 20 categories selected from Tan corpus which are to be used as testing data set consist of fortune, computer science, house, automobile, basketball, health, decoration and another 13 categories.

In the first three experiments, we select 20 categories that contain approximately the same number of documents from the 863 corpus (2003) as the training data set, and select the corresponding 20 categories from the 863 corpus (2004) as the testing data set. We also choose the 5 categories which also exist in the 863 corpus and 20 random categories from Tan corpus. In the last experiment, we use all categories in the 863 corpus as the training data set, and select the non-exist 20 categories in the 863 corpus from Tan corpus as the testing data set.

In the evaluation phase, we use MacroF1 and MicroF1 values to measure the performance of our classifier.

5.2 Test from the Same Source in Homogeneous Feature Space

We select the 863 corpus (2004) as the testing data set, and the twenty categories selected are the same as those selected from the 863 corpus (2003). The result of transfer learning is as shown in Table 1.

Table 1. Result of transfer learning from the same source corpus (%)

Testing data	MacroF1	MicroF1
863 corpus (2004)	73.0	71.2

Table 1 shows that MacroF1 and MicroF1 are to some extent lower than the best evaluation result in previous years, and it is due to the low value of "History and Geography", "Economy" and "Art" that directly influence the final result. After careful analysis of their knowledge bases, we find the reason is that the keywords selected from training documents cannot precisely represent their categories and cannot cover their whole keywords.

So, to examine the influence of different keyword selection to our categorizing result, we manually add some words into our list of stop words to make the selection of keywords more precise. We manually add some verbs, adjectives and adverbs such as "提出 'Propose', 好 'good', 高 'high', 表示 'express'. Then, we get a new transfer knowledge base, and do the categorization again. Table 2 shows how the categorization result changes after improving the precision of keywords in our experiment.

Table 2. Result after improving the precision on keyword selection (%)

Testing data	MacroF1	MicroF1
863 corpus (2004)	75.7	74.8

Although the precision and recall value of some categories decline slightly, Table 2 shows that the overall precise and recall value increase. Especially those whose precision and recall value are low before our changing increase dramatically, and the most prominent increase reaches about 28%. Also, MacroF1 and MicroF1 both increase about 3%, which almost reach the best evaluation result in previous years. From this result, we can see that if keywords are chosen properly, they can have a boosting effect on the efficacy of our transfer learning method.

5.3 Test from Different Source in Homogeneous Feature Space

In order to test the effectiveness of transfer knowledge and the proposed method in this paper, we use the corpus that are collected and organized by Dr. Songbo Tan in China Academy of Science Institute of Computing. Five categories in Tan corpus are the same or similar to the 863 corpus. Table 3 shows these five categories and the result of directly transferring the knowledge acquired in the 863 corpus (2003) to Tan corpus.

Table 3. Result from different source in homogeneous feature space (%)

Categories in the 863 corpus	Categories in the Tan Corpus	Precision	Recall	F1
Economy	Finance	84.7	99.3	91.4
Art	Aesthetics and Art	65.5	85.7	74.3
Astronomy, Earth Science	Astronomy	88.1	87.6	87.8
Literature	Literature and Art	94.2	74.5	83.2
Medicine, Health	Medicine	99	86.9	91.5
Macro-F1		86.5		
Micro-F1		88.1		

Table 3 shows that the MacroF1 and MicroF1 values of directly transferring knowledge acquired from 863 corpus to Tan corpus closely approximate the categorization result given by Dr. Songbo Tan, who conducted categorization tasks over Tan corpus in several traditional ways, which are shown in Table 4. However, the result of “Art” in 863 corpus (“Aesthetics and Art” in Tan corpus) and “Literature” in 863 corpus (“Literature and Art” in Tan corpus) is comparatively low. This is caused by the difference in the source of collecting documents in two corpuses. For example, “Aesthetics and Art” can be divided into “Aesthetics” and “Art”; and “Literature and Art” can be divided into “Literature” and “Art”. However, the overall result is satisfying.

5.4 Test of the Effectiveness of Transfer Knowledge Itself

To testify the effectiveness of the method of acquiring transfer knowledge, we choose Tan corpus to do cross validation. First, we still choose the previous 5 categories to do 3-cross validation, aiming to compare the result with the result shown in Table 3 in the previous section. Then, we randomly choose 20 categories in Tan corpus and do 3-cross validation, aiming to examine the effectiveness of our method at the macro level.

Firstly, we present the classification results done by Dr. Songbo Tan. He measured his corpus in five different ways including Centroid-Based, KNN, Winnow, Bayes and SVMTorch. The results are shown in Table 4.

Table 4. Baseline result provided by Dr. Songbo Tan (%)

Items	Centroid Based	KNN	Win-now	Bayes	SVM Torch
MacroF1	0.8632	0.8478	0.7587	0.8688	0.9172
MicroF1	0.9053	0.9035	0.8645	0.9157	0.9483

Then, we do 3-cross validation over previous 5 categories in Tan corpus. The result is as shown in Table 5.

Table 5. Test of 3-cross validation over 5 categories in Tan corpus (%)

Corpus	Macro-F1	Micro-F1
Tan corpus (5 categories)	96.1	97.5

By comparing the result shown in Table 3 and Table 5, we can see that the result of 3-cross validation over Tan corpus is 9~10% higher than transferring knowledge from 863 corpus to Tan corpus. And there are mainly two reasons for this discrepancy. One is caused by the difference in collecting documents in two corpuses. And the other is caused by the difference in the number of test documents.

Next, to further prove the effectiveness of our text categorization method, we randomly choose 20 categories in Tan corpus to do 3-cross validation. And the result is as shown in Table 6.

Table 6. Test of 3-cross validation over 20 categories in Tan corpus (%)

Corpus	Macro-F1	Micro-F1
Tan corpus (5 categories)	89.6	91.0

Table 6 shows that the Micro-F1 and Macro-F1 value of our method of acquiring transfer knowledge both reaches about 90% when it is applied to categorization tasks over one corpus. By comparison between Table 4 and Table 6, we can see that our result is satisfying and thus testify the efficacy of the transfer knowledge acquisition strategy proposed in this paper.

5.5 Weakly-Supervised Transfer Learning

To challenge the conventional assumption in machine learning that training data set and testing data set being in homogeneous feature space, and to testify the effectiveness of Algorithm 2, we acquire transfer knowledge from the 863 corpus (2003) and apply it to the 20 categories in Tan corpus but not in the 863 corpus. This is to construct a transfer knowledge base from unrelated domains so that we can meet the classification requirement in new text domains.

The 863 corpus and Tan corpus we use in this paper differ greatly in contents and publishing time, so we can regard the knowledge base we construct from the 863 corpus as an outdated knowledge base or cross-domain knowledge base. We firstly acquire transfer knowledge from the 863 corpus (2003) and construct a knowledge base according to Algorithm 1, and then extract 50 keywords from each category in Tan corpus by using methods like TF-IDF, and considering these keywords as the pre-knowledge for each category. Then, we form a new transfer knowledge base based on the previously constructed knowledge base. Finally, we build up our text classifier by extracting 50 keywords and their co-occurrence information from the testing data set according to Algorithm 4. Table 7 shows the result of transfer learning among heterogeneous feature spaces.

Table 7. Test of Transfer knowledge among heterogeneous feature spaces (%)

Testing data	MacroF1	MicroF1
Tan corpus	80.1	87.6

We can see from Table 7 that MicroF1 is comparatively higher. This means that the precision of our experiment is high on the whole. In contrast, the value of MacroF1 is comparatively low, which indicates that there is much difference in the value of precision and the value of recall. Also, we find that the precision of several categories, such as "Psychology", "Publication", "Job hunting", are much lower than others. After review their transfer knowledge base, we find that the keywords extracted by TF-IDF could not well represent their categories respectively; and that these three categories differ so greatly from categories in the 863 corpus that it is too difficult to extract transfer knowledge for them.

To further test the effectiveness of our transfer knowledge base, we increase the number of keywords extracted in Algorithm 2 from 50 to 100. Accordingly, when building up our text classifier in Algorithm 4, we also increase the number of keywords from 50 to 100, so that we can add more information into our transfer knowledge base. The result of transfer learning after the expansion of transfer knowledge base is shown as in Table 8.

Table 8. Test of transfer learning after increasing transfer knowledge base (%)

Testing data	MacroF1	MicroF1
Tan corpus	81.6	88.5

By comparing the result of Table 7 and Table 8, we can see that the value of precision of "Fortune" and "Publication" increases by approximately 10%, and other categories also increase in some degree. Although the values of MacroF1 and MicroF1 do not increase much, we can still draw the conclusion that the expansion of our transfer knowledge base has positive effect on the result of text categorization.

In order to further test the effect of expansion transfer knowledge base, we acquire transfer knowledge from both the 863 corpus (2003) and the 863 corpus (2004). The result of transfer learning after the expansion of transfer knowledge base is shown as in Table 9.

Table 9. Test of transfer learning after expanding transfer knowledge base (%)

Testing data	MacroF1	MicroF1
Tan corpus	81.4%	88.6%

By comparing TABLE 8 and TABLE 9, we find that although the value of precision and recall of each category changes in some degree, the value of MacroF1 and MicroF1 remains unchanged. This is because, firstly, there is not much difference between 863 corpus in the year of 2003 and 2004, so adding the 863 corpus (2004) does not add more useful features into our transfer knowledge base; and secondly, the knowledge extracted from the 863 corpus (2003) is already enough for us to obtain the linguistic information for our pre-knowledge, so adding the 863 corpus in the year of 2004 does not help much to our classification result. Still, the result of our experiment is satisfying. And this indicates that the transfer learning method proposed in this paper which aims at solving the difficulty in constructing cross-domain knowledge base is very effective.

6 Conclusions

In this paper, we present a novel strategy for acquiring transfer knowledge and apply it to automatic text categorization tasks among homogeneous and heterogeneous feature spaces. By conducting experiments across different corpuses and different domains, we get a satisfying outcome, which testifies the effectiveness of our method. However, in our methods, we only extract some basic linguistic information. So our future work may involve: (1) try to add more linguistic information to our method of acquiring transfer knowledge; (2) apply key technique in our method to public test set to further examine its efficacy. And actively explore other strategies for acquiring transfer knowledge as well as transfer learning methods.

Acknowledgments. This work is supported by the national natural science foundation of China (No. 61073130) and the project of National High Technology Research and Development Program of China (863 Program) (No. 2011AA01A207).

References

[1] Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 22(10), 1345–1359 (2010)

[2] Dai, W., Xue, G.-R., Yang, Q., Yu, Y.: Co-clustering based Classification for Out-of-domain Documents. In: Proceedings of the Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007), San Jose, California, USA, August 12-15, pp. 210–219 (2007)

[3] Xue, G.-R., Dai, W., Yang, Q., Yu, Y.: Topic-bridged PLSA for Cross-Domain Text Classification. In: Proceedings of the Thirty-first International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR 2008), Singapore, July 20-24, pp. 627–634 (2008)

[4] Ling, X., Dai, W., Xue, G.-R., Yang, Q., Yu, Y.: Spectral Domain-Transfer Learning. In: Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008), Las Vegas, Nevada, USA, August 24-27, pp. 488–496 (2008)

[5] Dai, W., Yang, Q., Xue, G.-R., Yu, Y.: Self-taught Clustering. In: Proceedings of the Twenty-Fifth International Conference on Machine Learning (ICML 2008), Helsinki, Finland, July 5-9, pp. 200–207 (2008)

[6] Dai, W., Chen, Y., Xue, G.-R., Yang, Q., Yu, Y.: Translated Learning: Transfer Learning across Different Feature Spaces. Advances in Neural Information Processing

[7] Ling, X., Xue, G.-R., Dai, W., Jiang, Y., Yang, Q., Yu, Y.: Can Chinese Web Pages be Classified with English Data Source? In: Proceedings the Seventeenth International World Wide Web Conference (WWW 2008), Beijing, China, April 21-25, pp. 969–978 (2008)

[8] Blitzer, J., McDonald, R., Pereira, F.: Domain adaptation with structural correspondence learning. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pp. 120–128 (2006)

[9] Salton, G., Buckley, C.: Term Weighting Approaches in Automatic Text Retrieval. Information Processing and Management 24(5), 513–523 (1988)

[10] Lewis, D.D.: Naïve(Bayes) at forty: The Independence Assumption in Information Retrieval. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 4–15. Springer, Heidelberg (1998)

[11] Yang, Y.M., Liu, X.: A Re-examination of Text Categorization Methods. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrival, Berkeley, CA, USA, pp. 42–49 (August 1999)

[12] Han, E., Karypis, G.: Centroid-Based Document Classification Analysis & Experimental Result. In: Zighed, D.A., Komorowski, J., Żytkow, J.M. (eds.) PKDD 2000. LNCS (LNAI), vol. 1910, pp. 424–431. Springer, Heidelberg (2000)

[13] Yang, Y.M.: An evaluation of statistical approaches to text categorization. Information Retrieval 1(1), 76–88 (1999)

[14] He, J., Tan, A.H., Tan, C.L.: A Comparative Study on Chinese Text Categorization Methods. In: PRICAL 2000 Workshop on Text and Web Mining, Melbourne, pp. 24–35 (August 2000)

[15] Nigam, K., Lafferty, J., McCallum, A.: Using maximum entropy for text classification. In: Proceedings of the IJCAI 1999 Workshop on Information Filtering, Stockholm, Sweden (1999)

[16] Wiener, E.: A neural network approach to topic spotting. In: Proceedings of the 4th Annual Symopsium on Document Analysis and Information Retrieval (SDAIR 1995), Las Vegas, NV (1995)

[17] Apte, C., Damerau, P., Weiss, S.: Text mining with decision rules and decision trees. In: Proceedings of the Conference on Automated Learning and Discovery Workshop 6: Learning from Text and the Web (1998)

[18] Lent, B., Swami, A., Widom, J.: Clustering association rules. In: Proceedings of the Thirteenth International Conference on Data Engineering (ICDE 1997), Birmingham, England (1997)

[19] Tan, S., Wang, Y.: Chinese Text Categorization Corpus-TanCorpV1.0., `http://www.searchforum.org.cn/tansongbo/corpus.html`

[20] Tan, S., et al.: A Novel Refinement Approach for Text Categorization. In: ACM CIKM (2005)