Chinese Patent Classification Based on Sense Disambiguation and Manifold Learning*

Lin FENG, Bo JIN*, Shilei SHAN

School of Computer Science and Technology, and School of Innovation Experiment, Dalian University of Technology, Dalian 116024, China

Abstract

Currently, patent data have gained increasing attention in the data mining area. However, the traditional method was set by many factors in patent classification, such as professional terms and high dimensions. In this article, we propose a word sense disambiguation model, which would effectively reduce the structure of the traditional machine word. The experimental result shows that the semantic disambiguation and feature reduction strategy can effectively improve the classification accuracy.

Keywords: Patent Classification; Semantic Disambiguation; Text Mining; Manifold Learning; Dimension Reduction

1 Introduction

The whole of text classification techniques include three aspects: text-based representation model, feature dimensionality reduction, and research of classification methods. The text-based representation model mainly studies the representation of the text, which items should be chosen, and the corresponding weight. The basic method is the vector space model (VSM), which is simplifying the text as “Bag of Words” (BOW) and carrying out the process of feature selection based on mathematical statistics and machine learning algorithms to predict the text class [1]. However, BOW is entirely constrained by the machine statistical approach and uses words as a feature item, ignoring the links among the semantic units. Conversely, the running speed of classification is influenced [2].

Many traditional algorithms are not able to deal with the text data, because of (1) The dimension of the sample is very high, and the utilization of Euclidean distance will fail [3]; (2) Bellman has pointed out the “curse of dimensionality problem” of the high-dimensional data [4].

In recent years, the introduction of the dimensionality reduction algorithm to text learning has gradually received more and more attention, such as the classic principal component analysis

*This study was supported by National Natural Science Foundation of P. R. China (Grant No. 61173163, 61173163), Natural Science Foundation of Liaoning Province, P. R. China (Grant No. 201102037), and Program for New Century Excellent Talents in University (Grant No. NCET-09-0251).

*Corresponding author.
Email address: jinbo@dlut.edu.cn (Bo JIN).
(PCA) algorithm [5], Multidimensional scaling transformation (MDS), LLE local neighbors, text manifold [6], Isomap [7] and the local keeping projection (LPP) [8]. Among of them, LPP extracts local information of data and can find the global linear mapping to achieve dimension reduction as PCA, obtaining the better learning results.

In this article, we apply the semantic disambiguation of the feature word to reflect the semantic concept. Then, through the manifold learning dimensionality reduction, we can extract the essential characteristics of the text vector and carry out classification training and class prediction more effectively.

2 Feature Sense Disambiguation

Word sense disambiguation is a process, which computers determine the words’ meaning according to their context. Chinese ambiguous words account for about 10% in the lexicon [9]. The paper has shown that the ratio can reach up to 42% in a large corpus. Word sense disambiguation is difficult and important in the natural language processing, which is an important procedure in machine translation, information retrieval, and other applications.

Along with machine learning development, many statistical methods have emerged for word sense disambiguation. Lu [10] pointed out that the key of the research in word sense disambiguation is the knowledge source. Recently, WordNet is commonly used as an English semantic lexicon. The Chinese semantic dictionary mainly contains HowNet and Tongyici Cilin. We used the Tongyici Cilin as the knowledge dictionary.

The statistical disambiguation methods mainly used statistical methods for automatic acquisition of knowledge. The commonly used statistical methods are decision trees, Bayesian model, neural network, support vector machine, maximum entropy, and vector space model (VSM). Zhang [11] compared and analyzed four statistical word sense disambiguation models and pointed out that the maximum entropy model and Bayesian model are better. A Chinese patent after the disambiguation processing case shown in Figure 1.

![Fig. 1: Example of feature disambiguation](image_url)

The upper part of the diagram is the original patent text content, and the lower part is the data organization form. The original words text is simply replaced with the word code combination. The synonym words are identical, and there are different codes for the polysemy.
3 Feature Dimension Reduction

3.1 LDA algorithm

Let dataset $X = [x_{11}, \ldots, x_{1N_1}, x_{21}, \ldots, x_{cN_c}] \in \mathbb{R}^{D \times N}$ contain $c$ classes, and the sample of each class $x_{ij} \in \Omega_i$, $i = 1, 2, \ldots, c$; the number of the $i$-th class is $N_i$. Assuming the PDF of the $i$-th class is $P(\Omega_i)$, the mean of each class is $\bar{x}_i = \sum_{x_{ij} \in \Omega_i} x_{ij}$, and the within class covariance matrix can be obtained: $V_i = \frac{1}{N_i} \sum_{x_{ij} \in \Omega_i} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^T$; further, the within class scatter matrix is $S_w = \sum_{i=1}^c p(\Omega_i) V_i$, and the between class matrix is $S_b = \sum_{i=1}^c p(\Omega_i) (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T$.

LDA is expected to find a linear projection $W = [w_1, w_2, \ldots, w_d] \in \mathbb{R}^{d \times D}$ in a low-dimensional space. The subspace $W$ satisfied within class scatter is minimum, and the between class scatter is maximum. Then, the optimization problem can be solved as follows:

$$
\begin{align*}
W_{opt} &= \arg \max_W \frac{Tr(W^T S_b W)}{Tr(W^T S_w W)}, \\
& \text{s.t. } W^T W = I
\end{align*}
$$

(1)

where $W^T W = I$ is the normalization constraint. The optimization problem can be transformed into a generalized eigen-decomposition problem: $S_b W_{opt} = \lambda S_w W_{opt}$. If the training samples of each class are not sufficient, and $S_w$ is not inversed, this leads to a small sample-size problem.

4 The Proposed Approach of Document Classification

In this article, we use the word sense disambiguation module to construct semantic features, and also use LDA to reduce the data dimension.

![Fig. 2: The process of text classification](image)

Step 1. Text preprocessing. We computed the word sense disambiguation, and then employed word sense to replace the word.
Step 2. Feature dimensional reduction and weight calculation. Five feature selection functions were employed, which are depicted in Table 1. Table 1 shows some typical methods in feature item selection and calculation.

Step 3. The classification model. The SVM model was used for text classification. It is trained by the training text. The SVM is tested with the text vector. The experiments are carried out to compare the results of word sense disambiguation and manifold dimension reduction.

<table>
<thead>
<tr>
<th>Table 1: Feature selection functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF-IDF</td>
</tr>
<tr>
<td>$d_{ik} = \sqrt{\sum_{k=1}^{n} \left(\frac{\log(f_{ik})+1}{\log(N) \cdot n_k}\right)^2}$</td>
</tr>
<tr>
<td>Chi Square</td>
</tr>
<tr>
<td>$\chi^2(t, c) = \frac{N(AD-CB)^2}{(A+C)(B+C)(A+B)(C+D)}$</td>
</tr>
<tr>
<td>Mutual Information (MI)</td>
</tr>
<tr>
<td>$I(t, c) \approx \frac{AN}{(A+C)(A+B)}$</td>
</tr>
<tr>
<td>Information Gain (IG)</td>
</tr>
<tr>
<td>$G(t) = -\sum_{i=1}^{m} P(c_i) \log(P(c_i)) + P(t) \sum_{i=1}^{m} P(c_i</td>
</tr>
<tr>
<td>$P(\bar{t}) \sum_{i=1}^{m} P(c_i</td>
</tr>
<tr>
<td>Document Frequency (DF)</td>
</tr>
<tr>
<td>$DF(t_k) = n_k$</td>
</tr>
</tbody>
</table>

5 Experiments and Analysis

The data are obtained from the China Patent Intellectual Property Office Website. The data include 4,000 Chinese patents, which are divided into eight categories as the standard of IPC. Each category contains 500 documents. The amount of data used in the experiments is provided in Table 2.

<table>
<thead>
<tr>
<th>Table 2: Patent data for test</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>class</strong></td>
</tr>
<tr>
<td>Training</td>
</tr>
<tr>
<td>Testing</td>
</tr>
</tbody>
</table>

C1 denote Human Necessities; C2 denote Transporting; C3 denote Chemistry; C4 denote Textiles; C5 denote Fixed Constructions; C6 denote Mechanic Engineering; C7 denote Physics; C8 denote Electricity.

In the experiment, the number of features was 16,069 without disambiguation processing, and after disambiguation, it was reduced to 1181. We selected 12 kinds (50 to 2,500) of feature numbers in the non-disambiguation classification experiments, and 8 kinds in the disambiguation experiments. The manifold dimensionality reduction was carried out on the basis of feature selection, and we used LDA as the dimensionality reduction algorithm, the target dimension is 10.
Table 3 records the situation of every classification with the increase of the number of features. From the table, we can see that classification results were improved along with the increase of the number of a feature, but when the feature number was too big, the situation tended to be stable. The more features, the more information, which is useful to the classification. However, with the increasing features, information quantity increases to a certain level, the amount of noise data also increases. Thus, the classification accuracy is eventually decreased in the trend. In addition, we can see that the accuracy rates of all kinds of classification are relatively different, and the reason will be the patent data itself category tagging errors or inappropriate classification granularity.

Table 4 shows the situation of classification after disambiguation, along with an increase in the number of features. From the data record, we can see that as the feature number increases, the accuracy rate of classification is improved, and tends to be stable. However, compared with the situation with disambiguation each kind of different characteristic numbers is increased by 5% to 15%. Without disambiguation categories, when the feature number up to 2500, the algorithm
efficiency will be very poor.

Figure 3 describes the results of different classification algorithms, including the original clas-
The classification accuracy after dimensionality reduction has a significant growth, which is about 26%. Meanwhile, the effect of dimensionality reduction is better than disambiguation; especially with the increase of the original characteristic, the manifold dimensional reduction method has a more prominent effect. When the characteristic is small, the effect of the word sense disambiguation process is superior to lower-dimensional manifold. The effect after hybrid disambiguation and dimensionality reduction is best. However, with the increase of characteristics, the manifold dimension’s reduction effect upgrade continues and it indicates its nature, which is that pre-screening features may affect its effect. When the characteristic is large, just the manifold dimension reduction can be used to obtain the best result, whereas a very large characteristic may lead to low efficiency.

When the characteristic is small, the classification result is best. However, when the characteristic is increased, the classification result has a sharp decline. After the word sense disambiguation, the total feature number is 1181; when the selected feature number is much more near the total number, then the text classification space has loopholes, and the manifold methods cannot process it. The total number of features without disambiguation is 16,069. When 2,500 features are selected, the manifold dimensionality reduction methods are effective.

6 Conclusion

This article analyzed the traditional classification model, and pointed out its problems. On the one hand, by performing semantic disambiguation on feature words, it replaces the word with the semantic features as feature items, and it reflects the semantics of the concept of the patent text. On the other hand, on the basis of analysis of the manifold methods, using manifold learning for dimensionality reduction on eigenvectors, it can extract the essential characteristics of the text vectors, and perform more effective classification training and category forecasts. Finally, using SVM classifiers to perform experiments, the results show that semantic disambiguation and manifold dimension reduction effectively improve the classification accuracy.

Acknowledgement

This work is supported by National Natural Science Foundation of P.R. China (Grant No. 61173163, 51105052), Program for New Century Excellent Talents in University (Grant No. NCET-09-0251), and Liaoning Provincial Natural Science Foundation of China (Grant No. 201102037).
References


