A Combined Weight method in Automatic Classification of Chinese Text

Shasha Liao  
Lab of Computational Linguistics, Dept. of Chinese Language, Creative Base of Cognitive Science, Tsinghua University, Beijing, 100084, China  
Email: liaoss03@mails.tsinghua.edu.cn

Minghu Jiang  
Lab of Computational Linguistics, Dept. of Chinese Language, Creative Base of Cognitive Science, Tsinghua University, Beijing, 100084, China  
Email: jiang.mh@tsinghua.edu.cn

Abstract - In this paper, we set a shielded level in a concept tree to use both the concept attributes from a semantic dictionary and the Chinese words to make the feature set. After comparing the weight theories and classification precise, of the eight methods, we give a new selection method, the CHI-MCOR weight method, which is derived from two normal methods which present well in our experiments. Our former experiment result shows that if we can set a proper shielded level, we can not only reduce the feature dimension but also improve the classification precise. The later result shows that the combined weight method makes a good balance between the fuzzy words which have a high occurrence and the dividing words which have a middle or low occurrence, and the classification precise is higher than any one of the weight methods.

I. INTRODUCTION

Automatic classification of Chinese text is a process which classifies the Chinese documents to several categories by their content. With the rapid development of the online information, automatic classification becomes one of the key techniques for handling and organizing the text data. It always has two main parts: the feature selection which reflects the documents to a feature vector space and weights the components of the vector, and the classifier, which classifies the documents to the right category by their feature vectors. Because the huge data of the document set and the hardness of reflecting the documents to feature vector, we need to construct a proper feature set[1]. Nowadays, the Chinese feature selection methods are mainly based on three forms: the word bag, the character bag and concept attribute. In the three forms, the word bag is the most popular one, because it is easy to be understood and handled. However, as the word is sometimes synonymous and dependent on its context, the word bag always ignores the relationship among the words and isolates all the words which are semantically related [3].

Because the concept space is much smaller than the word one, and the components are comparatively independent, the concept attributes are much better to reflect the content of the documents. Research shows that with the concept attributes and semantic analysis, we can get a much better vector space [2]. So we choose concept attribute as the main component of our feature selection method[3].

Feature weight method is also important in feature selection because it can not only give a proper threshold to reduce the feature dimension to a computable one but also strengthen the important features. There are two kinds of feature selection method: the first one is based on threshold filtering, including DF (document frequency), TF-IDF (Term Frequency–Inverse Document Frequency), MI (Mutual Information), CHI ($\chi^2$ statistics), IG (Information gain) and so on; and the second one is based on feature reconstruction, including Principle Component Analysis, Latent Semantic Indexing and so on. Because the first one is easier and faster,
II. THE CONCEPT EXTRACTION METHOD

The information of the concept extraction in this paper comes from HowNet[5], which is not only a semantic dictionary, but also a knowledge system referring to the concept of Chinese words and the relationship among them. So we use the DEF term of the Chinese word, which descript the word with defined concept attribute, to construct the feature reflection of the documents. In our method, we extract the concept attribute from the word as the reflection of the text, which will describe the internal concept information, and get the relationship among the words. Because there are 24,000 words in the HowNet and only nearly 1,500 concept attributes, the feature will be reduced to a stable dimensionality space with little information lose.

A. The Concept Weight Method based on HowNet

Because the different concept attributes have different expression powers, it is unwise to give every attribute the same weight. In fact, we need a strategy to give different expression powers to these attributes, so we consider two factors to weight them, one is the height of the weight node; the other is the number of the child nodes of the weight node. The height of the node is the most important factor because it shows the detail degree of the concept. Also, when a node have more child nodes, it means that in the cognize world, this concept is more complex and has more detail concept, and people would use its child nodes more and treat this concept as a more abstract one, so it should have a comparatively lower weight. Moreover, because the nine concept trees in HowNet are not equal, we give a different root weight to treat them differently.

At last, we give a weighting formula as follow:

$$W_{ik} = W_{tree_i} \cdot [\log((Droot_{ik} + 1)/2) + a + \frac{1}{Deep_k + b}]$$

In this formula, $W_{ik}$ is the weighting of node $k$ in tree $i$; $W_{tree_i}$ is the weighting of the tree root $i$, in case that there are nine trees in Hownet and they contribute differently in classification, and we give different weight to different trees; $Droot_{ik}$ is the distance between node $k$ and the tree root $i$; $CN_{ik}$ is the number of the child nodes of node $k$; $L$, $a$ and $b$ are the tempering factors, which are used to control the weighting range. According to the experiments, we set $L=0.15$, $a=1$ and $b=5$.

B. The Abstract Concept Attributes and the Shielded Level in the C-Tree

If the DEF term of a certain word contains only abstract concept, which has a weak expression power, it means that this DEF term does not describes the word precisely and the information gain is not enough. So we can not extract all the words into concept attributes. In this paper, we give a strategy to make a balance between the original words and the extracted concept attributes.

We use the concept tree of HowNet to calculate the expression power. By a selected shielded level, we divided these nodes into two parts, the strong ones and the weak ones. Because we mainly use the level of the node to decide its expression power, like formula 2, we set a level threshold, which is called shielded level. And for a word, if all the attributes in the DEF are above the shielded level, we consider that these attributes are weak in expression and give less information than the original word, and we do not extract this word into concept. The formula calculating the concept expression power of a word is as follow:

$$f(c) = \max_{j=0}^{m} k(c_j)$$

In this formula, $k(c_i)$ is the weight of attribute $i$ in the DEF term of word $c$; $m$ is the number of attributes in DEF term. This formula calculates all the attributes in a DEF term and decides whether the attribute or the word should be added to the feature set. If there is at least one attributes whose levels are higher than the shielded level, the expression power of the DEF terms are enough and we
added them into the feature set. Otherwise, the original word is added.

III. THE COMBINED WEIGHT METHOD

A. The Analysis of the Feature Set

When we extract the concept attribute to form the feature set, we convert a lot of words into the concept features, and get rid of the influence of the synonymy and dependence, which makes the classification precise much higher. However, because of the mass of weak concept and the words which are not in the HowNet, some Chinese words are given a comparatively lower weight and become the middle or low occurring feature. And there are still some specialty words and proprietary words which are only occur in one category and are not highly occurred in the whole documents and are very important in classification. Both of these words need a strategy to get a higher weight and contribute more in text classification and we analysis and experiment on the weighting methods which are widely used nowadays.

B. The Comparing Result of Seven Weight Method

We select seven common weight methods and test them, and focus mainly on their selection strategy and classification precise. The experiment gives us the following results:

From the analysis of the selected feature, we find that:

1. The DF, TF-IDF, CET (an improved method of IG), CDW and CHI methods prefer the high occurred words and they are greatly related. In our experiment, CHI is the best method in our experiments, which accords with the research of Yang Yiming[6]

2. MCOR method mainly choose the middle and low occurred feature, so its classification precise is low when the reduction rate is high. But with the increase of the feature dimension, its precise is increased highly and when the feature dimension is above 4000, its precise is higher than CDW, CET, DF, TF-IDF and MI methods.

3. MI method mainly selects the high and middle occurred feature, it can get a good classification precise but with the increase of the feature dimension, the precise is not improved visibly.

C. CHI Weight Method

The CHI weight method’s formula is as follow:

\[
\chi^2(t, c) = \frac{N \ast (AD - CB)^2}{(A + C)(B + D)(A + B)(C + D)}
\]

\[
\chi^2_{\max}(t) = \max_{i=1}^{m} \chi^2(t, c_i)
\]

In this formula, N is the total document number of the training set, c is a certain category, t is a certain feature, A is the number of the document which belong to category c and contain feature t, B is those which do not belong to category c but contain feature t, C is those which belong to category c but do not contain feature t, D is those which do not belong to category c and do not contain feature t.

CHI method is based on such hypothesis: if the feature is highly occurred in a specified category or highly occurred in other categories, it is useful in classification. Because CHI take the occurrence frequency into account, it prefers to select highly occurred words, and ignored the middle and low occurred words which maybe important in
D. MCOR Weight Method

The MCOR weight method’s formula is as follow[1]:

\[
MC - OR(t) = \sum_{i=1}^{m} P(C_i) \times |OR(t, C_i)| \\
= \sum_{i=1}^{m} P(C_i) \left| \log \frac{P(t|C_i)(1-P(t|C_{else}))}{(1-P(t|C_i)P(t|C_{else}))} \right|
\]

In this formula, \(P(C_i)\) is the occurrence probability of category \(C\), \(P(t|C_i)\) is the occurrence probability of the feature \(t\) when category \(C\) is occurred, \(P(t|C_{else})\) is the occurrence probability of the feature \(t\) when category \(C\) is not occurred. When \(P(t|C_i)\) is higher or \(P(t|C_{else})\) is lower, the weight of MCOR is higher. So, MCOR selects the features which are mainly occurred in one category and nearly not occurred in other categories. Because it does not consider the occurrence frequency of the feature, it prefer to select the words which are middle or low occurred in the document while highly occurred words are always occurred in more than one categories.

E. The Combined Weight Method

Because MCOR mainly selects the words whose occurrence frequencies are middle or low, its classification precise is low when the reduction is high. But with the increase of feature dimension, its precise is improved to a appreciable level. And CHI prefer to select the words whose occurrence frequencies are high, and it is one of the best feature selection methods [6]. As a result, when we combine this tow methods, we can make the advantages together and get a high classification precise[7]. So, we give a combined weight method based on CHI and MCOR:

\[
V(t) = \lambda V_{CHI}(t) + (1 - \lambda) V_{MCOR}(t) \quad 0 < \lambda < 1
\]

VCHI is the weight of feature \(t\) when we use CHI method, VMCOR is that when we use MCOR method. When we analysis the weigh given by these two methods, we find that the average weight of the features are different. For example, when the reduction is 50%, the range of the weight of CHI is (2.1,6.81), while that of MCOR is (1.15,1.76). Because CHI gives a much higher weight to all the features and its swing is wider, we should give a comparatively lower value to \(\lambda\). If not, the value depends too much on CHI and the combined weigh method is meaningless. So we need a proper value of \(\lambda\). In experience, we suppose that when the average weight of CHI and MCOR are the same, we can both get the advantage of the two and the classification precise will be the highest. So we think the best \(\lambda\) is as follow:

\[
\frac{\lambda}{1 - \lambda} = \frac{Mean(MCOR)}{Mean(CHI)}
\]

IV. EXPERIMENTS

This system is coded in Windows XP, and the coding tool is VS.Net. The corpus comes from the People Daily from 1996 to 1998. The corpus is unbalanced, and the training set is 1205, the test set is 755.

A. The Experiment Result of the Concept Extraction with Shielded Level

The experiment shows that with our concept extraction method, we can efficiently reduce the feature dimension and in this reduction, we do not lose useful information and the classification precise is much better because it filters the unnecessary noises.
Fig. 2. This is the classification precise of the system with different shielded levels. The y axis is the classification precise, and x axis is the categories of the classification and the last one is the average precise as the precise of the system.

From the result we conclude that only uses original words or concept attributes are both not very suitable. In the experiment, if we only use the concept attributes without any shielded levels, the precise is 90.9%, which is the lowest. And when we choose a proper level, for example, level 6, the precise is 93.7%, which is the highest.

Moreover, when we use concept attribute as the feature, the difference among different categories are less than that when we use word bags. This is probably because the feature selection based on original words depends much on the categories because if there are more special words in this field, it is easier to classify it from others. But when we use concept attributes, this difference between categories seems to be smaller and the curve seems to be much smooth.

**B. The Experiment Result of the CHI-MCOR**

In order to analysis the best value of $\lambda$, we vary $\lambda$ from 0 to 1.0. From the experiment, we found that when $\lambda$ is 0.3, the classification precise is the highest. This result accords to our hypothesis. Meanwhile, we find that when we use the combined weight method, the precise is always higher than other methods. For example, when $\lambda$ is 0 or 1, it is the precise of the MCOR method or CHI method. In our experiment, when $\lambda$ is 0.3, the precise is 94.0359%, which is 0.61% higher than CHI, 1.074% higher than MCOR.

Fig. 3. This is the average precise in CHI-MCOR. The y axis is the average precise, and x axis is the value of $\lambda$ in the formula which ranges from 0 to 1.

Below is the classification precise of CHI-MCOR, CHI and MCOR when $\lambda$ is 0.3.

Fig. 4. This is the precise of the six categories in three weight method, CHI, MCOR and CHI-MCOR. The y axis is the classification precise, and x axis is the categories of our test, the last one is the average precise.

Fig. 5. This is the average precise in CHI-MCOR. The y axis is the average precise, and x axis is the value of $\lambda$ in the formula which ranges from 0 to 1.

From the figure we can see that the combined weight method is much better in classification in politics category, it means that there are a lot of important words in politics category which are not highly occurred. So, when we use CHI-MCOR, its precise is 3.66% higher than we use CHI.
method. In fact, when we statistic the top ten of the occurred words in politics category, we find that they are not very high in the total statistics.

<table>
<thead>
<tr>
<th>NO.</th>
<th>Feature</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>政</td>
<td>5.7105</td>
</tr>
<tr>
<td>2</td>
<td>地方</td>
<td>5.0383</td>
</tr>
<tr>
<td>3</td>
<td>官</td>
<td>3.6262</td>
</tr>
<tr>
<td>4</td>
<td>国家</td>
<td>3.3958</td>
</tr>
<tr>
<td>5</td>
<td>商</td>
<td>2.7556</td>
</tr>
<tr>
<td>6</td>
<td>众</td>
<td>1.9959</td>
</tr>
<tr>
<td>7</td>
<td>干部</td>
<td>1.7792</td>
</tr>
<tr>
<td>8</td>
<td>吴</td>
<td>1.3168</td>
</tr>
<tr>
<td>9</td>
<td>市</td>
<td>1.2836</td>
</tr>
<tr>
<td>10</td>
<td>建立</td>
<td>1.2016</td>
</tr>
</tbody>
</table>

From the table we can see that the occurrence of the key feature in politics category is comparatively lower than other categories. For example, the frequency of the key word in economics category can reach to 49.7048. As a result, the classification precise of politics category is much higher when we use the CHI-MCOR weigh methods.

V. CONCLUSIONS

When we use concept as the feature of text classification, we can efficiently reduce the feature dimension and reflect the original feature space to a more stable one. By setting a shielded level, we can save the word whose DEF is weak in expression and avoid losing important information in concept extraction. When the shielded level is proper, the classification precise is much higher and more stable.

Because there are some dividing words which are not highly occurred but useful in text classification, we use CHI-MCOR method to combine two weight methods together. This method not only selects the highly occurred words, but also selects the dividing word whose occurrence frequency is middle or low. The experiment shows that it is much better than any one of the weight method.

ACKNOWLEDGEMENT

This work was supported by Tsinghua University 985 research fund; the excellent young teacher program, Ministry of Education, China; and State Key Lab of Pattern Recognition open fund, Chinese Academy of Sciences.

REFERENCES