Combining Statistical Information and Semantic Similarity for Short Text Feature Extension

Xiaohong Li, Yun Su, Huifang Ma, Lin Cao

Combining Statistical Information and Semantic Similarity for Short Text Feature Extension

Xiaohong Li[1], yun Su[1], Huifang Ma[1], lin Cao[*]

[1]College of Computer Science and Engineering,
Northwest Normal University Lanzhou Gansu,China
nwnulixiaohong@sina.com

Abstract. A short text feature extension method combining statistical information and semantic similarity is proposed，Firstly, After defining the contribution of word, mutual information, an associated word-pairs set is generated by comparing the value of mutual information with threshold , then it is taken as the query words set to search for HowNet. For each word-pairs, senses are found in knowledge base HowNet, and semantic similarity of query word-pairs are calculated. Common sememe satisfied condition is added into the original term vector as extended feature, otherwise, semantic relationship is computed and the corresponding sememe is expanded into feature set. The above process is repeated, an extended feature set is finally obtained. Experimental results show the effectiveness of our method.

Keywords: short text, statistical correlation, semantic similarity, hownet, feature extension

1 Introduction

With the explosion of the network new media and online communication, short texts in diverse forms such as news titles, micro-blogs, instant messages, have become the main stream of information exchange. Most of the traditional classification methods are not good at short text classification and failed to accomplish the task effectively. Therefore, how to improve the efficiency of classifying the mass of short text has become the researching focus.

Recently, new classifying methods on short text appeared. Kim[1] proposed a novel language independent semantic (LIS) kernel, which is able to effectively compute the similarity between short text documents. Wang[2] presented a new method to tackle data sparseness problem by building a strong feature thesaurus (SFT) based on latent Dirichlet allocation (LDA) and information gain (IG) models. Methods mentioned above are mainly pays more attention to the concept and the correlation of texts to obtain the logic structure. Therefore, their classifying performance has been improved a little. Yuan[3] presented a short text feature extension method based on frequent term sets, larger search space of algorithm result in higher time complexity, particularly,

adfa, p. 1, 2011.

when the scale of the background knowledge increased, the dimension of feature word set would increase dramatically.

A short text feature extension method combining statistical information and semantic similarity was proposed to overcome the drawbacks of the above. The flowchart is shown in figure1.

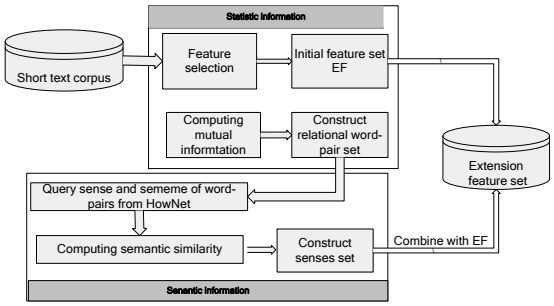

Fig. 1. Clustering algorithm flow

2 Preliminary Knowledge

In this section, we briefly introduce some related knowledge from two aspects: contribution of words and mutual information.

Contribution of Words

We define the contribution[4] of words as:

$$contr(w,d) = \frac{f(w,d)}{f_{max}(d)} \tag{1}$$

where f(*w*, *d*) represents the number of the word *w* in document *d*, $f_{max}(d)$ is the maximum number of word occurred in document d.

Thus, the contribution of the word w to the class C_k can be defined as the sum of the contribution of the word w to all documents in C_k, which is computed as follow:

$$CONTR(w,C_k) = \sum_{j=1}^{N_k} contr(w,d_j) \tag{2}$$

When *k* takes different value, the CONTR(*w*, C_k) denotes the contribution of the same characteristic towards to different category.

Mutual Information

Let $T=\{w_1,w_2\}$denote a word-pairs, we can compute mutual information[5] between the word-pairs *T* and the class *C* according to the following formula:

$$MI(T,C) = H(C) - H(C \mid T) \tag{3}$$

Where H(*C*) is the entropy of whole classification system *C*, H(*C*|*T*) is the conditional entropy of *C* given a word-pairs *T*.

$$H(C) = -\sum_{k=1}^{K} p(C_k)\log_2 p(C_k) \tag{4}$$

$$\begin{aligned} H(C \mid T) = &-p(T)\sum_{k=1}^{K} p(C_k \mid T)\log_2 p(C_k \mid T) \\ &- p(\bar{T})\sum_{k=1}^{K} p(C_k \mid \bar{T})\log_2 p(C_k \mid \bar{T}) \end{aligned} \tag{5}$$

3 Feature extension algorithm and weight computing

Semantic Similarity in HowNet

HowNet[6] is a common sense knowledge database that reveals the relationship between concepts as well as concepts and attributes.

Suppose there are two words w_1 and w_2, *m and n* is the number of senses of w_1 and w_2 respectively. We describe this using the following formula: $S_1 = \{ s_{11}, s_{12},\ldots,s_{1n}\}$, $S_2 = \{ s_{21},s_{22},\ldots,s_{2m}\}$. Word similarity[7] of w_1 and w_2 is the maximum senses similarity of s_{1i} and s_{2j}:

$$ss(w_1, w_2) = \max_{i=1\ldots n,\, j=1\ldots m} sim(s_{1i}, s_{2j}) \tag{6}$$

It has been concluded that if $ss(w_1,w_2) > \beta$, *CS* symbolized the intersection of S_1 and S_2, *CS* is not empty. The model is shown in figure 2:

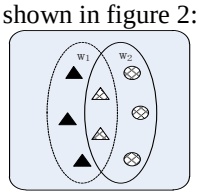

Fig. 2. Sense relationship of word-pairs

White circle denotes the senses of w1, black triangle represents the senses of w2, while triangle is common senses of w_1 and w_2.

Feature Extension Algorithm

The goal of expanding the short text feature set is to describe the topic and content of texts as accurate as possible. A new method identified as FEASS (feature extension algorithm based on semantic similarity) has been proposed in this paper aimed at the above principle.

```
Program: FEASS (F, α, β, δ)
input: candidate set F, three thresholds
output: expanded feature set EF
1: initial: EF = Φ, U =Φ, k=1,
2: while (F≠Φ)
      Calculate CONTR according to formula (2);
      If CONTR(w, Ck)> α, then EF= EF∪{w};
4: for(i=1; i<|EF|; i++)
for(j=i+1; j<|EF|; j++)
          if MI((wi,wj),Ci)>β   U = U∪ {(wi,wj)};
5: while k<|U|
5.1 compute ss(wi,wj) according to formula (6);
5.2 if ss(wi, wj) > δ   then EF = EF∪CS
    else: EF = EF∪(sik∪ sjh);
5.3: k++, goto 5.1
6: return EF
```

4 Experiment Results and Analysis

We conduct three experiments on SVM classifier to evaluate our method, experimental setup and results are described in detail in the following subsections.

Dataset

The experimental data in this paper comes from the China Knowledge Resource Integrated Database (CNKI), we collect 35603 piece of article published from during the period from 2013 to 2015. At last, we keep two thirds pieces of article title for each class as training samples and leave the remaining one third pieces of article title in total as test samples.

Experiment and Analysis

The Influence of Different Parameters.
We carried experiments on SVM when the parameters take different values for the parameter α, β and δ, and choose several representative results to be shown in Table 1.

The results on both classifier are the best while α=0.10, β=0.05, δ=0.25. Classification performance is very poor when the values of α and β is small. There are two reasons for this phenomenon, one is that redundant features have not been screened out, the other is that extending some boring words into features set. Conversely, classification efficiency appear to decline when the values of two parameters is great.

Table 1. The influence of different threshold on classifier

parameters			*SVM*		
α	β	δ	P	R	F1
0.05	0.05	0.20	68.35	65.38	66.83
0.05	0.10	0.20	69.83	70.24	70.03
0.10	0.05	0.25	75.48	73.75	74.25
0.10	0.10	0.25	70.35	61.26	65.49
0.15	0.20	0.30	68.27	59.38	63.59
0.15	0.25	0.30	63.23	57.41	60.23

The Efficiency of Classification before and after Feature Expanding.
Fig.3 shows results of our method before and after feature extension on SVM classifier. We can find that our method achieves 4.24%, 2.90%, 3.39%, 7.78%, 5.85%, 7.17%, 8.66%, 3.23%, 6.06% and 4.58% improvements with F-measure for Finance, Geology, Oceanography, Math, Astronomy, Agriculture, Biology, Physics, Medicalscience, and Computer respectively. Good results of Precision are achieved as well.

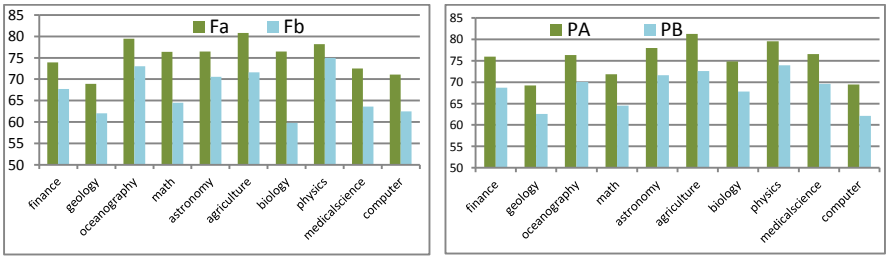

Fig. 3. F-measure and precision on SVM before and after feature extension

Comparison of Different Feature Extension Algorithm.
In this part, we compare the performance of FEASS with FEMFTS[8] (Feature Extension Method using Frequent Term Sets) and SCTCEFE[9], (Short Text Classification Considering Effective Feature Expansion), they are all state-of-the-art short text feature extension approach.

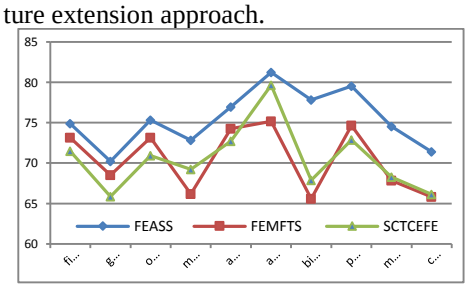

Fig. 4. Precision on SVM

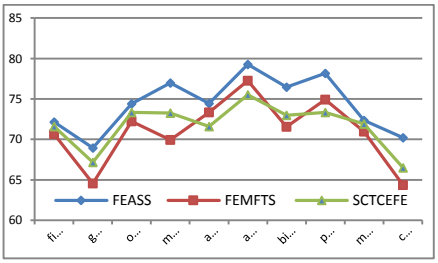

Fig. 5. F-measure on SVM

It can be seen from fig.4 to fig.5 that Precision and F-measure of FEASS algorithms on SVM classifier. The best F-measure reached 84.52, Precision achieved 83.67 respectively, so our algorithm is slightly higher than FEMFTS and SCTCEFE.

5 Conclusion

In this paper, we propose a feature set extension algorithm for short text classification. We find our method can achieve a good performance, so the feature compensatory method is feasible with the aid of external knowledge base. In the future, we plan to resolve how to find the expansion of the 'key' information in the corpus, and add as little noise as possible in features set to achieve the goal of effective 'extension'.

6 Acknowledgments

This work was supported in part by the Natural Science Foundation for Young Scientists of Gansu Province, China(Grant No. 145RJYA259, 1606RJYA269), Project of Gansu Province Department of Education (No. 2015A-008), Key Project of NWNU (No. NWNU-LKQN-14-5).

7 References

1. Kim Kwanho, Chung Beom-suk, Choi Yerim et al. "Language independent semantic kernels for short-textclassification," *Expert Systems with Applications*, Vol.41(2), pp. 735-743, 2012.
2. Wang, BingKun, Huang YongFeng, Yang WanXia and Li, Xing, "Short text classification based on strong feature thesaurus," *Journal of Zhejiang University SCIENCE C*, Vol. 13(9), pp. 649-659, 2012.
3. Yuan Man, Feature Extension for Short Text Categorization Using Frequent Term Sets[J]. Procedia Computer Science 31. 2014: 663 – 670.
4. Chen YuZhong, Fang MingYue, Guo WenZhong, Reasearch on Multi-Label Propagation Clustering Method for Microblog Hot Topic Detection[J]. Pattern Recognition and Machine Learning, Vol. 28, No. 1, 2015.
5. L Batina, B Gierlichs, E Prouff,M Rivain etal. Mutual Information Analysis: a Comprehensive Study[J] Journal of Cryptology, 2011, 24(2):269-291
6. Q. Liu, S.J.li, "Word's semantic similarity computation Based on the HowNet", The 3rd Chinese lexical and semantic proseminar, Taipei, China, 2002.
7. Pan LiQiang, Zhang Pu, Xiong AnPing, Semantic Similarity Calculation of Chinese Word[J]. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 5, No. 8, 2014
8. Peat H J, Willet P. The Limitations of Term Co-occurrence Data for Query Expansion in Document Retrieval Systems[J]. Journal of American Society for Information Science, 1991, 42(5): 378～ 383.
9. Liu MingXuan , Fan XingHua , A Method for Chinese Short Text Classification Considering Effective Feature Expansion[J].International Journal of Advanced Research in Artificial Intelligence, Vol. 1, No. 1, 2012.