A Text Classifier Based on Sentence Category VSM

ZHANG Yun-liang[1, 2] ZHANG Quan[1]

1. Institute of Acoustics, Chinese Academy of Sciences, Beijing 100080, China
2. Graduate School of the Chinese Academy of Sciences, Beijing 100039, China
zylioa@126.com, zhq@mail.ioa.ac.cn

Abstract. VSM is a mature model of text representation for categorization. Words are commonly used as dimensions of feature space of VSM, but words only provide little semantic information. Sentence category theory is an important component of HNC theory and can provide abundant information about meaning, structure and style of a sentence. We use sentence categories as dimensions of feature space, reduce the dimensionality by dividing mixed sentence categories and reform the weights by tfc-weighting algorithm. By simple vector distance calculation, we can get the parameters of the classifier and execute the categorization. The average precision and recall of our classifier are acceptable and can be improved by other HNC techniques.

Key words: Text categorization；HNC theory；Nature Language Processing; Sentence category；Computer application

1 Introduction

With the rapid increase of electronic documents, especially the growth of the internet and intranet, automatic text processing becomes more and more important. Text categorization is usually the basic and important components of text processing applications. It will benefit other applications, such as IR, Q&A.

Statistical categorization is the major method to resolve the categorization problem. Vector Space Model (VSM) was created by G. Salton in 1960s and now widely used in text representation [1]. All categorization methods based on VSM commonly include feature vector generation, dimensionality reduction of feature space, machine learning and categorization execution [2-4].

The idea of VSM representation is to discretize a consecutive text and form a vector in particular feature space. Words are the most common features of VSM and there are also models that take phrases, terms or Chinese character as features [5]. The weight of every element of a vector is usually the number of occurrences or its transform. Tfc-weighting is a transform algorithm.

Typically, the dimensionality of the feature space is more than tens of thousands. It is too high to result in too much calculation time and memory space. So dimensionality reduction is necessary. Feature selection and feature merger are two useful approaches. Feature selection is to discard some features non-informative or non-dipartite. Document Frequency Threshold, Information Gain, CHI, Mutual Information are means of feature selection [6-10]. Feature merger is an approach to compress the feature space. Latent Semantic Indexing (LSI) and using Ontology such as WordNet, HowNet, thesaurus and HNC concept primitive woods are all feature merger [11-12]. With machine learning methods, we can get the parameters of the classifier and decide an incoming document belong to which category. Simple vector distance, naïve Bayes, k-Nearest Neighbor (k-NN), Support Vector Machine (SVM), Artificial Neural Network (ANN), Decision Tree, Voting and so on are all applied methods [13].

2 Related HNC knowledge

Chinese characters, words, phrases, terms are common features of VSM. With the increase of the feature granularity, the ambiguity decrease. But the dimension will increase because combination and the appearance times of every feature of the vector decrease. To build better classifier, we should use more semantic features, but sentence level semantic is restricted because syntax analysis of sentences mainly provide structure information and not compatible for text categorization. To use sentence information as features of a VSM, we must resolve two problems. The first one is how to represent the sentence with simple symbol with enough semantic information. The second one is how to merge the billions of dissimilar sentences and receive acceptable probability of every feature of the vector.

Hierarchical Networks of Concepts (HNC) theory is a theory about NLP created by Huang Zengyang. HNC theory derives from the primitivity of Chinese characters, Chomsky's Universal Grammar, Quillian's Semantic Networks and Fillmore's Case Grammar [14]. Primitivity is a treasure of Chinese characters. Most of Chinese characters have powerful combination ability and the meanings of characters in the combination are basically derivable. Chinese characters are materials to build the concept primitive woods and other infrastructures. From the other three sources, Huang Zengyang takes the hierarchical and network structure and some useful segments. Sentence category theory is one of the most important components of HNC theory. Huang have found that there are limited categories of sentences. It is supposed to fit all languages in the world and now verified in Chinese and English with large-scale corpus.

There are 57 groups of basic sentence categories. Sentence categories in the same group have almost same meanings but a little different in form. The 57 groups of sentence categories are divided into 307 species. And there are also mixed sentence categories, a mixed sentence category contain at least two categories not in format but connotation [15]. A sentence category, whether basic or mixed, is made up of semantic chunks. Semantic chunk is a junior segment of sentence. It may be a word, a phrase, even a degenerated sentence. Semantic chunks are divided into main semantic chunks and auxiliary semantic chunks according to its importance to the expression of sentence category. Main semantic chunk is the essential component of sentence category, though perhaps some main semantic chunks are omitted, but they must be implied somewhere. An auxiliary semantic chunk is dispensable and hard to be speculated by context. A sentence category corresponds with an ordinal series of main semantic chunks.Eq.1 is an example of sentence category.

$$T3=TA+T+TB+ [\#T3C\#] \quad (1)$$

In this example, T3 is the name and the symbol of the sentence category. It represents a sentence prototype which is used to express an information transmission activity. TA is the chunk that represents the agent of the transmission. T is chunk that represents the activity of the transmission .TB is the chunk that represents the object of the transmission. TC is the chunk that represents the content of the transmission. In Eq.1 TC is packaged in brackets and pound signs, which means that TC is a junior sentence or a phrase that reformed from a junior sentence .Eq.2 is an example of mixed sentence category. It is mixed by T3 and Y30.

$$T3Y30*32=TA+T3Y30+TB+YC \quad (2)$$

Sentence categories include a lot of syntax, semantic and pragmatic information, so it is more informative than distinct words or phrases. Sentence categories are so abstractive and primitive that different sentences in superficial layer may have same sentence category in deep. Sentence.3 and Sentence.4 both belong to T3 sentence category, though they have different words and phrases.

I have told him that I will be back. (**3**)

Jim teaches me how to make a report. (**4**)

HNC theory and corresponding techniques lead to a new text categorization approach that used the sentence categories as dimensions of VSM feature space.

3 Sentence category VSM

Sentence category VSM uses sentence categories instead of words as dimensions of feature space. We should analyze a document first as usual but with different tools [16-17]. With the work of HNC sentence categories analysis tools, a text was changed into a series of symbols. And then we separate different sentence categories and count the number of occurrences of each sentence category in the document.

The feature space has high dimensionality. There are 307 different basic sentence categories and 93942(307×306) different mixed sentence categories that mixed by two basic sentence categories. So theoretically, there are at least 94249 different categories. Each category corresponds with one dimension in the feature space. If we consider mixed sentence category mixed by 3 or more basic sentence categories and the levels of a sentence category in context, the dimensionality will increase hundreds of folds. So dimensionality reduction is necessary. We adopt two hypotheses. First, we ignore the differences of the same sentence category in different levels, that is, sentence categories in sentences and degeneration sentences are regarded as equal. And if it is a mixed category, we divided it into different basic sentence categories with equal probability. For example, if a sentence has the sentence category T3Y30*32, we will plus the feature vector dimension T3 sentence category 0.5 and Y30 sentence category 0.5. After this processing step, we will get a feature vector V as Eq.5.

$$V_{dj}=(SC_{1j},SC_{2j},\cdots\cdots,SC_{nj}) \quad (\mathbf{5})$$

SC_{ij} $i \in N \quad 1 \le i \le n$ represents the no.i sentence category of the vector derived from document j.

Considering taking into account the frequency of the word throughout all documents in the collection, we reform the feature vector with tfc-weighting, namely TF-IDF weighting and normalization of the vector. The formula is as Eq.6.

$$a_{ij} = \frac{\log(TF_{ij}+1.0)*\log(N/DF_i)}{\sqrt{\sum_k [\log(TF_{kj}+1.0)*\log(N/DF_k)]^2}} \quad (\mathbf{6})$$

Where a_{ij} is weight of dimension i of the vector that represent document j. TF_{ij} is the number of occurrences of sentence category i in document j. N is the number of the documents in the collections. DF_i is the number of documents which include sentence category i. A document d_j can represent by a vector $\overrightarrow{d_j}$ as Eq.7.

$$\overrightarrow{d_j} = (a_{1j}, a_{2j}, \ldots\ldots, a_{nj}) \quad (\mathbf{7})$$

4 Training & categorization

To give the classifier usable parameters, training is necessary. We use simple vector distance calculation. In this algorithm, we use the vectors of classes to calculate the arithmetic mean of every

class and acquire a central vector which can represent the class by and large. The training algorithm is as follows:

Step1: Sentence category analyses of the documents in training collection.

Step2: Get vectors of every document.

Step3: Use arithmetic mean to calculate the central vectors of every class.

When new documents incoming, we also analyze them and receive results of sentence category analysis. And then the results will be changed into a vector in sentence category feature space. Then we calculate the vector distance with every central vector. The algorithm to classify documents is as follows:

Step1: Sentence category analyses of the documents.

Step2: Change the documents which need to be classified into vectors.

Step3: Calculation the similarity between incoming vectors and class central vectors use formula as Eq.8.

$$\mathrm{sim}(\vec{d}_i, \vec{d}_j) = \frac{\sum_{k=1}^{n} a_{ik} \times a_{jk}}{\sqrt{(\sum_{k=1}^{n} a_{ik}^2)(\sum_{k=1}^{n} a_{jk}^2)}} \quad (8)$$

Where $\vec{d}_i$ is a central vector of class *i* and $\vec{d}_j$ is a vector which represent incoming document *j*.

5 Performance evaluation

We use a collection of 1610 documents to test our categorization method. The documents are all news materials from internet and classified into 15 categories as politics system, politics activity, economy, education, science and technology, art, military affairs, health, environment, religion, living, laws, disasters and others.

We use recall and precision to evaluate the performance [18-19]. Precision of a particular category C_i (represented by precision (C_i)) is percentage of the number of documents classified correctly in Ci (represented by N_{cci}) divided by the number of documents classified into category C_i (represented by N_{cti}). Recall of a particular category C_i (represented by recall (C_i))is percentage of the number of documents classified correctly in Ci (represented by N_{cci})divided by the number of the documents should be classified into category C_i (represented by N_{tci}). To evaluate the classifier, we use average precision and average recall, and they are average precision and recall of all categories. The formulae are as Eq.9 to Eq.14. Eq.11 and Eq.12 are macro averaging formulae; Eq.13 and Eq.14 are micro averaging formulae.

$$precision(c_i) = \frac{N_{cci}}{N_{cti}} \quad (9)$$

$$recall(c_i) = \frac{N_{cci}}{N_{tci}} \quad (10)$$

$$precision_{oa} = \frac{1}{m}\sum_{i=1}^{m} precision(c_i) \quad (11)$$

$$recall_{oa} = \frac{1}{m}\sum_{i=1}^{m} recall(c_i) \quad (12)$$

$$precision_{ia} = \frac{\sum_{i=1}^{m} N_{cci}}{\sum_{i=1}^{m} N_{cti}} \quad (13)$$

$$recall_{ia} = \frac{\sum_{i=1}^{m} N_{cci}}{\sum_{i=1}^{m} N_{tci}} \quad (14)$$

We have both close test and open test. In the open test, we take 80% documents as train set and the other 20% as test set. Since a document often relates to two categories or more and belong to a most correlative category, we use two standards to evaluate our method.

$Precision_{oa}$, $recall_{oa}$ and $Precision_{ia}$, $recall_{ia}$ are average performance indexes. The performance is listed in table1. The performance is a middle level comparing with other VSM models, but in this classifier, we use sentence category only. If we use domain analysis and concept primitive woods, the performance will be improved. And it can also be improved by better MT methods.

Table 1. The performance of the classifier

	Close test	Open test
$Pricision_{oa}$	86.58%	72.99%
$Recall_{oa}$	88.27%	74.31%
$Pricision_{ia}$	86.87%	75.63%
$Recall_{ia}$	89.15%	74.91%

6 Conclusion

VSM is a mature text representation method and usually the dimensions of feature space are Chinese characters, words, phrases, etc. These features provide little semantic information. As is known, text is a whole and these features deviate from the original text very much. HNC theory and sentence category with corresponding analysis techniques provide a new approach to count more semantic information.

We use sentence categories as the dimensions of the feature space, and split the mixed sentence category to reduce the dimensionality. The vector dimensions are converted to equivalent the number of occurrences of different basic sentence categories in a particular document. We use simple vector distance to get central vector of every class and calculate the similarity between a coming document vector and category central vectors. Use average precision and recall of one and three related categories return, we evaluate the performance. The results are acceptable and can be improved on with domain analysis of HNC theory.

References

1. G.Salton and M.E.Lesk. Computer evaluation of indexing and text processing. Journal of the ACM,15(1): 8-36, January 1968
2. Aas K, Eikvil A. Text categorization: a survey, Norwegian computing center technical report, June 1999
3. Tang Yi-fang, Niu Li, et al. Automated text categorization. Journal of Guangxi Normal University , 19(4): 50-55, December 2001
4. Pang Jianfeng, Bu Dongbo, et al. Research and implementation of text categorization system based on VSM. Application Research of Computers, 18(9): 23-26, September 2001
5. Wang Mengyun, Gao Suqing. The System for Automatic Text categorization Based on Chinese Character Vector. Journal of the China Society for Scientific and Technical Information, 19(6): 644-649, December 2000
6. Yiming Yang, Jan O. Pedersen. **A** Comparative Study on Feature Selection in Text categorization. Proceedings of the Fourteenth International Conference on Machine Learning : 412-420, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 1997
7. Zhou Qian, Zhao Ming-sheng, et al. Study on feature selection in Chinese text categorization. Journal of Chinese Information Processing 18(3): 17-23, May 2004
8. Chen Xuetian, Li Ronglu. Using maximum entropy model for text categorization. Computer Engineering and Applications, 40(35): 78-79,195, December 2004
9. Huang Hai-ying, Lin Shi-min, et al. A study of text categorization on Concept Space. Computer Science, 30(3): 46-49, March 2003
10. Wang Min-chun, Wang Zheng-ou, et al. Rough set text categorization rule extraction based on CHI value. Computer Applications, 25(5):1026-1028,1033, May 2005
11. Wang Tianjiang, Ye Weiguo. Text categorization based on integrating LSI with k-nearest neighbor. J. Huazhong University of Sci. & Tech. (Nature Science Edition), 32(4): 59-60, 86. April 2004
12. Shi Yong-feng, Zhao Yan-ping. Comparison of text categorization algorithms. Wuhan University Journal of Nature Sciences, 9(5): 798-804, May 2004
13. Zhang Jian, Li Chunping. WordNet-based Concept Vector Space Model for Text Categorization. Computer Engineering and Applications, 42(4): 174-178, February 2006
14. Huang Zengyang. HNC Theory. Beijing: Tsinghua University Press, 1998
15. Miao Chuan-jiang. Studies on the Knowledge of sentence category in HNC theory. Beijing: Institute of Acoustics, Chinese Academy of Sciences, August 2001
16. Wei Xiang-feng. The software platform for expanded sentence category analysis based on the HNC theory. Beijing: Institute of Acoustics, Chinese Academy of Sciences, May 2005
17. Jin Yao-hong. Language processing techniques and applications based HNC theory. Beijing: Sciences Publication Inc, April 2006
18. Cheng Ze-kai, Lin Shi-min. Methods on accuracy evaluation of Text Classifier. Journal of the China Society for Scientific and Technical Information, 23(5): 631-636, October 2004
19. SONG Fengxi, GAO Lin. Performance evaluation Metric for text classifiers. Computer Engineering, 30(13) : 107-109,127, July 2004