A CUSTOMIZABLE TEXT CLASSIFIER FOR TEXT MINING

Yun-liang Zhang[1*] and Quan Zhang[2]

*[*1]Institute of Acoustics, Graduate School, Chinese Academy of Sciences, Beijing 100080, China*
E-mail: yunlzhang@mails.gucas.ac.cn
[2]Institute of Acoustics of the Chinese Academy of Sciences, Beijing 100080, China
E-mail: zhq@mail.ioa.ac.cn

ABSTRACT

Text mining deals with complex and unstructured texts. Usually a particular collection of texts that is specified to one or more domains is necessary. We have developed a customizable text classifier for users to mine the collection automatically. It derives from the sentence category of the HNC theory and corresponding techniques. It can start with a few texts, and it can adjust automatically or be adjusted by user. The user can also control the number of domains chosen and decide the standard with which to choose the texts based on demand and abundance of materials. The performance of the classifier varies with the user's choice.

Keywords: Text mining, Text categorization, Nature Language Processing (NPL)

1 INTRODUCTION

Text mining is a method of obtaining new information from a large collection of texts. It is similar to data mining because both try to find some useful unknown information. The resource for data mining is the database in which the data is structured, but the resource for text mining is unstructured (Hearst, 1999). Usually the data in a database have some consanguineous relations. Texts in a large corpus or on the Internet, however, are very different. Therefore, the first step in text mining is to get a proper collection. A user perhaps does not know what the knowledge is but perhaps knows that the knowledge is related to some entities, and the texts should be written in a particular style. Text mining can be done by hand, but it is better done by computer. Therefore, we have developed a customizable text classifier based on sentence categorization of the Vector Space Model.

2 VECTOR SPACE MODEL

Statistical categorization is the major method to resolve the categorization problem. The Vector Space Model (VSM) was created by G. Salton in the 1960s and is now widely used in text representation (Salton & Lesk, 1968). All categorization methods based on VSM commonly include feature vector generation, dimensionality reduction of feature space, machine learning, and categorization execution (Aas & Eikvil, 1999; Tang, Liu, et al., 2001; Pang, Bu, et al., 2001).

The idea of VSM representation is to discretize consecutive text and form a vector in a particular feature space. Words are the most common features of VSM, but there are also models that take phrases, terms, or Chinese characters as features (Wang & Gao, 2000). 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 takes up considerable calculation time and memory space. Therefore, dimensionality reduction is necessary. Feature selection and feature merger are two useful approaches. Feature selection discards non-informative or non-dipartite features. Document Frequency Threshold, Information Gain, CHI, and Mutual Information are means of feature selection (Yang & Peterson, 1997; Zhou, Zhao, et al., 2004; Chen & Li, 2004; Huang, Lin, et al., 2003; Wang, Wang, et al., 2005). Feature merger is an approach to compress the feature space. Latent Semantic Indexing (LSI) and use of ontologies such as WordNet, HowNet, thesaurus, and HNC concept primitive words are all feature mergers (Wang & Ye, 2004; Shi & Zhao, 2004). With machine learning methods, we can get the parameters of a classifier and decide to which category an incoming document belongs. 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 (Zhang & Li, 2006).

3 RELATED HNC KNOWLEDGE

Chinese characters, words, phrases, and terms are common features of VSM. With the increase of feature granularity, ambiguity decreases; however, the dimensions increase because the combination and appearance occurrences of every feature of the vector decrease. To build a better classifier, we must use more semantic features, but sentence level semantic parsing is restricted because syntactic analysis of sentences mainly provides structural information and is not compatible with text categorization. To use sentence information as a feature of a VSM, we must resolve two problems. The first is how to represent a sentence yet retain simple symbols with enough semantic information. The second is how to merge the billions of dissimilar sentences and receive an acceptable probability of every feature in the vector.

Hierarchical Networks of Concepts (HNC) theory is a theory about Nature Language Processing created by Huang Zengyang. HNC theory derives from the primitivism of Chinese characters, Chomsky's Universal Grammar, Quillian's Semantic Networks, and Fillmore's Case Grammar (Huang, 1998). Primitivism is a treasure of Chinese characters. Most Chinese characters have powerful combinatorial ability, and the meanings of characters in combinations are basically derivable. Chinese characters are materials with which to build the concept primitive words and other infrastructures. From the other three sources, Huang Zengyang (1998) takes the hierarchical and network structure combined with other useful segments. Sentence category theory is one of the most important components of HNC theory. Huang has found that there are a limited number of categories of sentences. His theory supposedly fits all languages in the world and has now been verified in Chinese and English with a large-scale corpus.

There are 57 groups of basic sentence categories. Sentence categories in the same group have almost same meanings but with little differences in form. The 57 groups of sentence categories are divided into 307 species. There are also mixed sentence categories - a mixed sentence category contains at least two categories not in the same format but with the same connotation (Miao, 2001). A sentence category, whether basic or mixed, is made up of semantic chunks. A semantic chunk is a junior segment of sentence. It may be a word, a phrase, or even a degenerated sentence. Semantic chunks are divided into main semantic chunks and auxiliary semantic chunks according to their importance to the expression of the sentence category. A main semantic chunk is the essential component of a sentence category. Although some main semantic chunks may be omitted, they must be implied somewhere. An auxiliary semantic chunk is dispensable and difficult to be speculated about by context. A sentence category corresponds with an ordinal series of main semantic chunks. Equation 1 is an example of sentence category.

T31=TA+T + [#T3C#] (1)

In this example, T31 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 transmission. T is chunk that represents the activity of the transmission. TC is the chunk that represents the content of the transmission. In Equation 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. Equation 2 is an example of a mixed sentence category. It is mixed by T3 and Y30.

T3Y30*32=TA+T3Y30+TB+YC (2)

Sentence categories include syntax, as well as semantic and pragmatic information, so they are more informative than distinct words or phrases. Sentence categories are so abstract and primitive that different sentences in a superficial layer may have the same sentence category in depth. Sentence 3 and Sentence 4 both belong to the T31 sentence category although they have different words and phrases.

李校长 || 宣布 || [# 傅老师 || 重返 || 学校#]。
Li xiaozhang xuanbu fu laoshi chongfan xuexiao. (3)
President Li declared that Professor Fu will be back to school.
T31J[#T3C#]=[#T2b0J#]

周济 || 指出 || [# 教学评估 || 是 || \{ 提高 | 教育质量 }的关键 / #]。
Zhouji zhichu jiaoxue pinggu shi tigao jiaoyu zhiliang de guanjian. (4)
Zhouji points out that teaching evaluation is the key to increase the education quality.
T31J[#T3C#]= [#jDJ#DC=\{!31XY401*211J}/#]

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

4 SENTENCE CATEGORY VSM

Sentence category VSM uses sentence categories instead of words as dimensions of feature space. Using it, we first analyze a document as usual but with different tools (Wei, 2005; Jin, 2006). With the work of HNC sentence categories analysis tools, a text is changed into a series of symbols. Finally, 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(307x306) different mixed sentence categories that are formed by mixing two basic sentence categories. Therefore, theoretically, there are at least 94249 different categories. Each category corresponds to one dimension in the feature space. If we consider a mixed sentence category consisting of three or more basic sentence categories and the levels of a sentence category in context, the dimensionality will increase a thousand fold. Therefore, 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 of sentences and degenerate sentences are regarded as equal. If a category is mixed, we divide it into different basic sentence categories with equal probabilities. For example, if a sentence has the sentence category T3Y30*32, we will add the feature vector dimension T3 sentence category 0.5 and Y30 sentence category 0.5. After this processing step, we arrive at a feature vector V as shown in Equation 5.

$$Vd_j=(SC1_j,SC2_j,\cdots\cdots,SCn_j) \qquad (5)$$

SCij $i \in N$ $1 \le i \le n$ represents the i^{th} sentence category of the vector derived from document *j*.

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 shown in Equation 6.

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

where a_{ij} is the 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, and DF_i is the number of documents that include sentence category *i*. A document d_j can be represented by a vector $\vec{d}_j$ as shown in Equation 7.

$$\vec{d}_j = (a_{1j}, a_{2j}, \ldots\ldots, a_{nj}) \qquad (7)$$

5 RESULTS

We used the entities user concerns as the filter and then processed the selected texts. Usually recall and precision are used to evaluate the performance (Cheng & Lin, 2004; Song & Gao, 2004). The precision of a particular category C_i (represented by precision (C_i)) is the percentage of the number of documents classified correctly in C_i (represented by N_{cci}) divided by the number of documents classified into category C_i (represented by N_{cti}). The recall of a particular category C_i (represented by recall (C_i)) is the percentage of the number of documents classified correctly in C_i (represented by N_{cci}) divided by the number of the documents that should be classified into category C_i (represented by N_{tci}). To evaluate the classifier, we use average precision and average recall, which are the average precision and recall of all categories. The formulae are shown in Equations 8 to 13. Equations 10 and 11 are macro averaging formulae; Equations 12 and 13 are micro averaging formulae.

$$precision(c_i) = \frac{N_{cci}}{N_{cti}} \qquad (8)$$

$$recall(c_i) = \frac{N_{cci}}{N_{tci}} \tag{9}$$

$$precision_{oa} = \frac{1}{m}\sum_{i=1}^{m} precision(c_i) \tag{10}$$

$$recall_{oa} = \frac{1}{m}\sum_{i=1}^{m} recall(c_i) \tag{11}$$

$$precision_{ia} = \frac{\sum_{i=1}^{m} N_{cci}}{\sum_{i=1}^{m} N_{cti}} \tag{12}$$

$$recall_{ia} = \frac{\sum_{i=1}^{m} N_{cci}}{\sum_{i=1}^{m} N_{tci}} \tag{13}$$

Our experiment was based on 10 collections. Every collection had a seed of 10 texts, 50 related texts and 50 unrelated texts, both types with focused entities. The performance is listed in Table 1.

Table 1. Performance of the classifier

AVERAGE PRECISION		AVERAGE RECALL	
MACRO	MICRO	MACRO	MICRO
0.729	0.756	0.743	0.749

6 REFERENCES

Aas, K. & Eikvil, A. (1999) *Text categorization: a survey*. Norwegian computing center technical report.

Chen, X. & Li, R. (2004) Using maximum entropy model for text categorization. *Computer Engineering and Applications 40(35)*: 78-79,195.

Cheng, Z. & Lin, S. (2004) Methods on Accuracy Evaluation of Text Classifier. *Journal of the China Society for Scientific and Technical Information 23(5):* 631-636.

Hearst, M. (1999) Untangling Text Data Mining. *Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland*.

Huang, H., Lin, S., et al. (2003) A study of text categorization on Concept Space. *Computer Science 30(3):* 46-49.

Huang, Z. (1998) *HNC Theory*. Beijing: Tsinghua University Press.

Jin, Y. (2006) *Language Processing Techniques and Applications Based on HNC Theory*. Beijing: Sciences Publication Inc

Miao, C. (2001) *Studies on the Knowledge of Sentence Category in HNC Theory*. Beijing: Institute of Acoustics, Chinese Academy of Sciences.

Pang, J., Bu, D., et al. (2001) Research and implementation of text categorization system based on VSM. *Application Research of Computers 18(9)*: 23-26.

Salton, G. & Lesk, M. (1968) Computer evaluation of indexing and text processing. *Journal of the ACM 15(1):* 8-36.

Shi, Y. & Zhao, Y. (2004) Comparison of text categorization algorithms. *Wuhan University Journal of Nature Sciences 9(5):* 798-804.

Song, F. & Gao, L. (2004) Performance evaluation Metric for text classifiers. *Computer Engineering 30(13):* 107-109,127.

Tang, Y., Niu, L., et al. (2001) Automated text categorization. *Journal of Guangxi Normal University 19(4):* 50-55.

Wang, M. & Gao, S. (2000) 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.

Wang, M., Wang, Z., et al. (2005) Rough set text categorization rule extraction based on CHI value. *Computer Applications 25(5):1026-1028,1033*

Wang, T. & Ye, W. (2004) Text categorization based on integrating LSI with k-nearest neighbor. *J. Huazhong University of Sci. & Tech. (Nature Science Edition) 32(4)*: 59-60, 86.

Wei, X. (2005) *The Software Platform for Expanded Sentence Category Analysis Based on the HNC Theory*. Beijing: Institute of Acoustics, Chinese Academy of Sciences.

Yang, Y. & Peterson, J. (1997) 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.

Zhang, J. & Li, C. (2006) WordNet-based Concept Vector Space Model for Text Categorization. *Computer Engineering and Applications 42(4)*: 174-178.

Zhou, Q., Zhao, M., et al. (2004) Study on feature selection in Chinese text categorization. *Journal of Chinese Information Processing 18(3)*: 17-23.