Text Categorization for Authorship based on the Features of Lingual Conceptual Expression[*]

Quan Zhang [a], Yun-liang Zhang [a,b], and Yi Yuan [a]

[a] The Institute of Acoustics, CAS, Beijing 100080, P. R. China
[b] Institute of Scientific & Technical Information of China, Beijing 100038, P. R. China

{zhq, yuan}@mail.ioa.ac.cn

Abstract. The text categorization is an important field for the automatic text information processing. Moreover, the authorship identification of a text can be treated as a special text categorization. This paper adopts the conceptual primitives' expression based on the Hierarchical Network of Concepts (HNC) theory, which can describe the words meaning in hierarchical symbols, in order to avoid the sparse data shortcoming that is aroused by the natural language surface features in text categorization. The KNN algorithm is used as computing classification element. Then, the experiment has been done on the Chinese text authorship identification. The experiment result gives out that the processing mode that is put forward in this paper achieves high correct rate, so it is feasible for the text authorship identification.

Keywords: Lingual Conceptual Space, Hierarchical Network of Concepts (HNC) theory, Text Categorization, KNN Algorithm, Authorship of a text.

1. Introduction

Along with the progress of social information industry, especially with the Internet development, more and more documents exist in the electronical form. It provides convenience for information automatic processing. The text categorization is the basic work for the automatic processing, and it is the foundation for the information retrieval, information mining, and question-answer system too.

In some applications, it is need to identify the text authorship. The identification need use the documents written by the author as reference. Since the literary style of an author is relatively steady in some time, if we can mine the style character of the author then the identification is realized. Of course, the more documents which are gathered, the better result which is achieved. In many cases, the text authorship identification is treated as text categorization as the researchers can accomplish the work according the text categorization way. In this way, the frequency of glossary, punctuation, n-Gram string, syntax feature, average sentence length, the length of paragraph, and so on, are be used as the features to identify the author of a text

[*] The paper is supported by the National Fundamental Research Project (973 Project) (Grant No. 2004CB318104) and The Knowledge Innovation Engineering Project of The Institute of Acoustics, CAS (Grant No. 0654091431).

[Burrows J. F. 1987, Baayen R. H. 1996, David I Holmes 1997, Yuta Tsuboi 2002]. The texts are often literature, and sometimes they are internet text, such as E-mail [Olivier de Vel 2001].

Presently, there are a few research works in the Chinese text authorship identification, and many of them orient to the linguistic research for the concrete literature. Sun and Jin [Yi-jian Jin, Xiao-ming Sun, and Shao-ping Ma, 2003] use the vector space model (VSM) which takes the syntax structural words as feature to identify the text authorship, they achieved good result in novel author identification. The best precision of pattern matching, KNN algorithm and SVM algorithm are 89.51%, 91.54%, and 93.58% separately. Ma and Chang study the E-mail authorship identification successively [Jian-bin Ma 2004, Shu-hui Chang 2005]. In addition, Wu introduces the HowNet knowledge base; he performs the text authorship identification according to the evaluation method of glossary semantic similarity. His best marco-average F-measure is 86.23% that is achieved in 202 People Daily texts written by 5 reporters [Xiao-chun Wu 2006].

In text categorization, one important aspect is text expression. The text is mainly expressed as discrete glossary. As to Chinese, it is also expressed as Chinese character, phrase, term, n-gram string, etc. When the discrete unites are obtained, they can be used in feature vector constructing.

These features that come from natural language directly can represent the topic of the text and the writing style of the author in some aspects. However, these features also arouse the spare data and too large feature space in the computing, because there are more than ten thousand basic words in natural language, and the total of words even come to million, the dimension of the corresponding vector space is huge. In order to perform the computing, it is necessary to reduce the dimension.

Huang put forwards the linguistic conceptual space according his Hierarchical Network of Concepts (HNC, in short) theory [Zengyang Huang 1998, 2004]. The HNC conceptual primitives of the linguistic conceptual space and its word knowledge expression can be used to reduce the dimension. In fact, the structure of HNC conceptual primitives is a tree list, which the nodes are semantic primitives. The word meaning is expressed with the primitives, and it can accomplish the computing of correlation in the glossary, finding the thesaurus, and reducing the dimension.

Therefore, we explore the Chinese text authorship identification with the KNN classifier that adopts the HNC conceptual primitives as the features carrier. We explain the HNC conceptual primitives and the semantic expression in the following firstly.

2. The HNC conceptual primitives and the semantic expression

The network of primitives provides a necessary semantic description system in HNC, which expresses the semantic with formalization and conceptualization. The basic unite of the primitives network is conceptual tree, the node in the tree is the element for semantic expression, the relation between trees and the nodes in one tree provides the relevancy of the primitives. There are 456 trees, which are defined in HNC.

Fig. 1 is an example for the conceptual tree. This branch provides the definition and reference for its dominative category. It defines part concepts of the professional activities in detail. The letter and the string, such as a, a1, a119, are the concept identifier for the node, and the content in the brackets is the comment for the conceptual node. All these are merely abstract definition, not for the concrete instance. When expressing the concrete instance, we need to combine more concepts according its meaning. For example, the “U.S.A. Government” can be expressed in “a119+fpj2*304”, the meaning of the “a119” can be found in the Fig. 1, and the “fpj2*304” refers in particular to the United Stats.

The goal of the HNC conceptual primitives is to take out and generalize the commonness of the things, and identify them with a well-defined string.

The network just describes the meaning of the words. As to the denotative function of a word, the HNC introduces five elements for the abstract concepts: v (verb), g (noun), u (attribute), z (value), and r (effect); two elements for the concrete concepts: w (matter) and p (people); and one element between the abstract and concrete concept: x (quality). These elements are called “classification of concept”. They are also primitives, and need to combine to express a word denotative function.

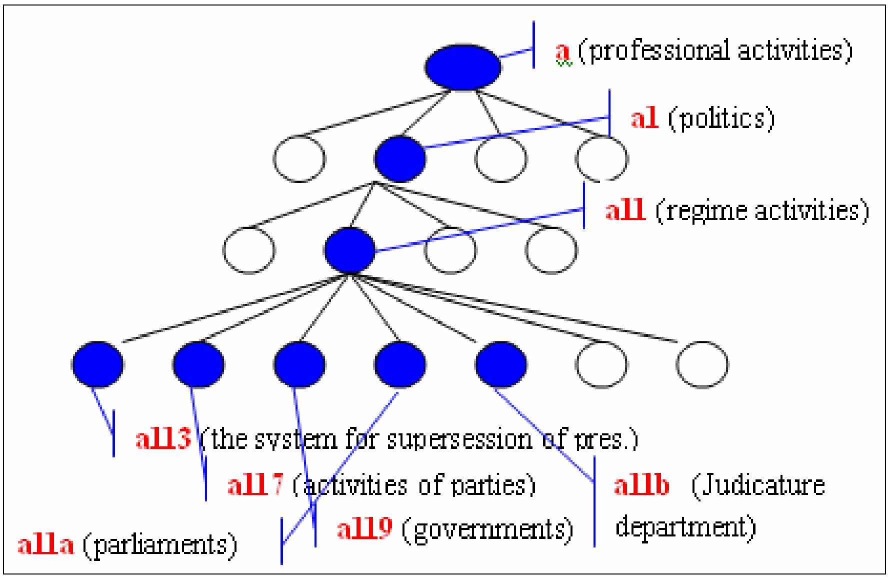
a (professional activities)
a1 (politics)
a11 (regime activities)
a113 (the system for supersession of pres.)
a117 (activities of parties)
a11b (Judicature department)
a119 (governments)
a11a (parliaments)

Figure 1: A branch of the conceptual tree.

Table 1: Chinese word in HNC primitives expression.

Word	Classification	Conceptual meaing
哀愁 Sad and worry	v,g	v7132+v7200e72
巡警 Cop	p	pa41+va11
诱发 Arouse	v	((va71,v9441)#(v900;v80))

We will adopt the classification of concept as categorization feature. The basic condition of the categorization feature is that the concepts which belong to one classification must not be too few, and the distinguish ability between the classifications is enough. Then we use the 12 kinds classification as feature.

Table 2: The 12 kinds classification for the feature.

No.	Name	The symbol of the classification of concept
1	Abstract noun	g/r/ xr / gr
2	Concrete noun	p/w
3	Entity	f
4	Verb	v/vv
5	Attribute and quality	u/ug/gu/x/ jx/px /gx / rx
6	Adverb	uu/uv

7	**Quantifier**	**zz/zzv**
8	**Time**	**j1**
9	**Space**	**j2**
10	**Syntax structural words**	**l0-l5/l8**
11	**conjunction**	**l6-l7/l9-lb**
12	**Idiom**	**f1-fb**

3. The algorithm and evaluation [Yun-liang Zhang 2006]

The effect of the KNN（K-Nearest Neighbor, KNN）algorithm is better in all of the VSM algorithm, and it is easy to implement, and its adaptability is robust. The basic idea of this algorithm is that the K texts in the training set which are the nearest to the input text are gotten, and then determine the input text type according the K texts.

The Formula 1 shows how to compute the similarity between the texts, where the a_{ik} indicates the k dimension value of the feature vector d_i

$$\mathrm{sim}(\vec{d}_{\mathrm{i}}, \vec{d}_{\mathrm{j}}) = \frac{\sum_{\mathrm{k=1}}^{\mathrm{n}} a_{\mathrm{ik}} \times a_{\mathrm{jk}}}{\sqrt{(\sum_{\mathrm{k=1}}^{\mathrm{n}} a_{\mathrm{ik}}^2)(\sum_{\mathrm{k=1}}^{\mathrm{n}} a_{\mathrm{jk}}^2)}} \tag{1}$$

The a_{ik} can be calculated as Formula 2.

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

Where the TF_{ij} is the frequency of the feature *i* occurred in the document *j*， *N* is the total of texts in the text set, DF_i is the document frequency of the feature *i*.

Then, the Formula 3 shows how to calculate the probability that represents the input text to belong to which type in the text set.

$$p(\vec{x}, C_j) = \sum_{\vec{d}_i \in KNN} sim(\vec{d}_x, \vec{d}_i) p(\vec{d}_i, C_j) \tag{3}$$

Where $p(\vec{d}_i, C_j)$ is categorization function, i.e. if $\vec{d}_i$ belongs to C_j ，then the value is 1, less the value is 0.

We categorize the input text to the type with the largest probability.

We evaluate the categorization result with MAFM（Micro-Average F-Measure）[Yun-liang Zhang 2006].

4. The experiment and result

The texts in this paper are novels that are download form Internet. There are 25 authors' novels; the detail is shown in Table 3. We sample 20 texts for each author as testing text, and divide into two set equally, one for train and the other for testing.

Table 3: The experiment texts

Author	Ba Jin(巴金)	Ding Ling (丁玲)	Jin Yong (金庸)	Lao She (老舍)	Liang Yun-sheng (梁羽生)	Lu Xun (鲁迅)	Lu Yao (路遥)
Amount	148	58	1618	415	730	32	280
Author	Ma Feng (马烽)	Mao Dun (茅盾)	Qian Zhong-shu（钱钟书）	Qiong Yao (琼瑶)	Qu Bo (曲波)	Shen Cong-wen (沈从文)	Cao Xun-qin (曹雪芹)
Amount	54	19	75	950	32	72	80
Author	Wang Shuo (王朔)	Wang Xiao-bo 王小波	Yang Muo (杨沫)	Ye Sheng-tao (叶圣陶)	Yu Hua (余华)	Zhang Ailing (张爱玲)	Zhang Henshui (张恨水)
Amount	175	76	74	30	105	59	425
Author	Zhao Shu-li (赵树理)	Zhou Li-bo (周立波)	Zhou Mei-sen (周梅森)	Luo Guang-bin (罗广斌)		Amount	
Amount	56	51	336	30		5916	

Table 4: The result of Abstract noun feature

K	1	2	3	4	5	6	7	8	9	10	11
MAFM	0.871	0.871	0.863	0.88	0.871	0.88	0.896	0.888	**0.9**	0.896	0.892
K	12	13	14	15	16	17	18	19	20	21	22
MAFM	0.892	0.884	0.884	0.888	0.888	0.888	0.884	0.884	0.888	0.888	0.884

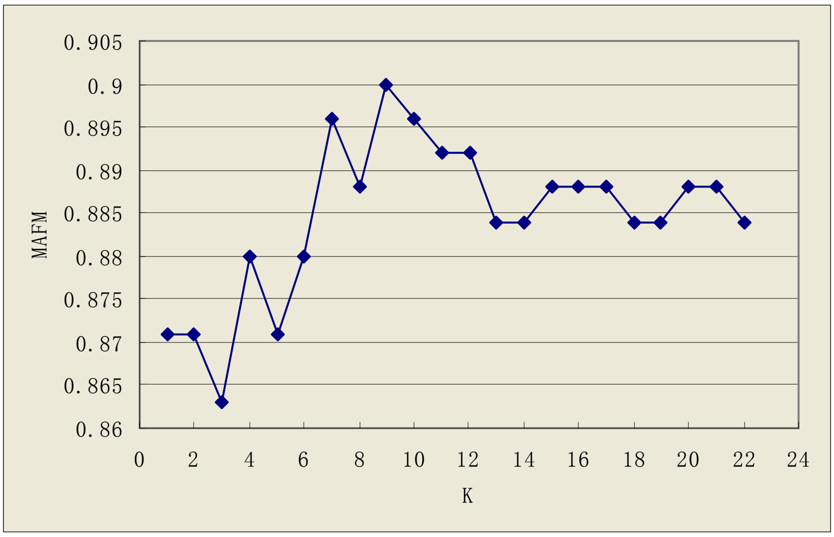

Figure 2: The result of Abstract noun feature

Table 5: The result of Concrete noun feature

K	1	2	3	4	5	6	7	8	9	10	11
MAFM	0.606	0.606	0.586	**0.618**	0.59	0.586	0.574	0.538	0.546	0.554	0.558
K	12	13	14	15	16	17	18	19	20	21	22
MAFM	0.554	0.554	0.55	0.526	0.518	0.518	0.522	0.506	0.514	0.506	0.502

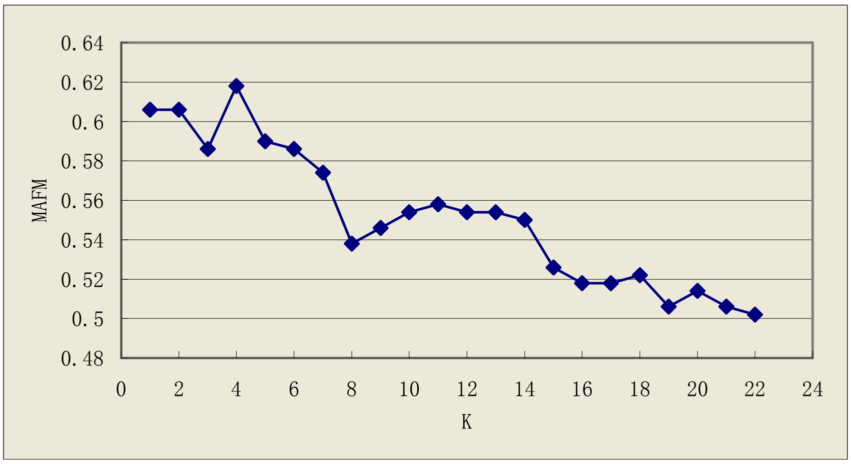

Figure 3: The result of Concrete noun feature

Table 6: The results of 12 kinds features

The Classification of the concepts	MAFMmax	The best K
Abstract noun	0.9	9
Verb	0.88	6
Adverb	0.819	4
Attribute and quality	0.731	10 or 11 or 14 or 15
Entity	0.655	8
Syntax structural words	0.655	4
Concrete noun	0.618	4
Time	0.582	10
Quantifier	0.546	27
conjunction	0.534	14 or 15
Space	0.526	4 or 7
Idiom	0.502	8

5. Discussion

We adopt the simple KNN algorithm as the computing tool, and the HNC conceptual primitives as the features carrier that can express the word meaning, then we contracture a mode to identify the text authorship. The experiment result indicates that this mode can work well, and achieves high correct rate. It means that the HNC conceptual primitives are suitable to express the text feature in text authorship identification.

Meanwhile, comparing with the natural language words as text feature, there are also some advantages when take the HNC conceptual primitives to express the text feature:

1. As the HNC conceptual primitives give the words meaning in concept, so some words with different form can be expressed in same symbol when they have the same meaning, such as the “distress” and “sad” they have the same symbol in HNC conceptual primitives. Then, they can be treated as same one.

2. The conceptual primitives are hierarchical symbols; they are suitable to compute the correlation. It is easy to reduce the dimension by merging the low frequency concept into a high frequency concept, according the correlation among concepts.
3. The HNC conceptual primitives are a kind of artificial semantic symbol, their meaning is well defined, and the natural language words always have ambiguous meaning. It is hard to get the clear meaning, and the HNC conceptual primitives provide the convenience: we can obtain the word meaning in the HNC conceptual primitives form after sentence analysis according the HNC processing.

References

Baayen R. H., H. van Halteren, and Tweedie F J. 1996. Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3):121-131.

Burrows J. F.. 1987. Word patterns and story shapes: the statistical analysis of narrative style.*Literary and Linguistic Computing,* 2:61-70.

David I Holmes. 1997. Stylometry:its origins, development and aspirations. *Joint international conference of the association for computers and the humanities and the association for literary and linguistic computing*. Norway: University of Bergen, 98-103.

Jian-bin Ma. 2004. The Study on the Authorship Mining for Chinese E-mail documents based on SVM, *Master's Dissertation*, Hebei Agriculture University.

Olivier de Vel, A. Anderson, M. Corney, and G. Mohay. 2001. Mining E-mail Content for Author Identification Forensics. *SIGMOD:Special Section on Data Mining for Intrusion Detection and Threat Analysis*, 55-64.

Shu-hui Chang. 2005. The Study on the Authorship Identification for Chinese E-mail documents Based on the Literary Style, *Master's Dissertation,* Hebei Agriculture University.

Xiao-chun Wu, Xuan-jing Huang, and Li-de Wu. 2006. Authorship Identification Based on Semantic Analysis. *Journal of Chinese Information Processing*, 20(6):61-68.

Yi-jian Jin, Xiao-ming Sun, and Shao-ping Ma. 2003. Stylistics based Writer Identify on Internet. *Journal of Guangxi Normal University (Natuarl Science Edition)*, 21(3):62-66.

Yun-liang Zhang, and Quan Zhang. 2006. A Text Classifier Based on Sentence Category VSM. *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, 244~249. Wuhan.

Yuta Tsuboi. 2002. Authorship Identification for Heterogeneous Documents. *Master's Thesis,* Nara Institute of Science and Technology.

Zengyang Huang. 1998. *The Hierarchical Network of Concepts theory.* Tsinghua University Press, Beijing .

Zengyang Huang. 2004. *The Fundamental theorem and mathematic physics expression of the language concept space*. Ocean Press, Beijing.