Short Text Feature Selection for Micro-blog Mining

Zitao Liu
International School of Software
Wuhan University
Wuhan, China
zitao.whu@gmail.com

Wenchao Yu
International School of Software
Wuhan University
Wuhan, China
issyuwenchao@gmail.com

Wei Chen
School of Electronics & Information Engineering
Sichuan University
Chengdu, China
hugePuff@gmail.com

Shuran Wang
School of Electronics & Information Engineering
Sichuan University
Chengdu, China
sran0124@gmail.com

Fengyi Wu
School of Electronics & Information Engineering
Sichuan University
Chengdu, China
fying.wood@gmail.com

***Abstract*—Feather selection is a process that extracts a number of feature subsets which are the most representative of the original meaning from original feature set. It greatly reduces the text processing time and increases the accuracy because of removing some data outliers. With the rapid development of Web 2.0 and the further evolution of the Internet, short text like micro-blog plays an important role in people's daily life. However, existing feature selection methods cannot effectively extract these short text features, and greatly reduce the classification and clustering performance of short text. In this regard, we propose a novel feature selection method based on part-of-speech and HowNet. According to the composition of the text property, we choose the words with larger amount of information by different part-of-speech, and then expand the semantic features of these words based on HowNet, in this way the short text has more useful features. We use test data set collected from sina micro-blog and adopt the micro average and macro average of F1-Measure to evaluate the effects of short text classification. The results show that the short text feature selected by our method has a good amount of information, as well as good classification results.**

Keywords- short text; feature selection; HowNet; part-of-speech

I. INTRODUCTION

With the increasing growth of the Internet, a large number of short text data expand in a geometric rate. How to improve the efficiency of accessing the integrated information resource through information fusion has become the research focus [1][2]. Take micro-blog for example, micro-blog is a new multimedia mini blog that allows users to publish short text messages for all or a limited group to read through SMS, instant messaging devices, email and other forms[3]. In recent years, micro-blog swept the world at an alarming rate, and have produced large amount of short text with short length and different data structure, which contains a lot of useful information [4][5]. However, similar to micro-blog, the reply message of a forum, feedbacks, SMS, instant messaging and e-mail have less content ranged from a few dozen words to a hundred words or so. Compared to the feature selection technology in the long text, the main problem in short text feather selection is: in short text, the feature space is sparse and difficult to fully exploit the correlation between their features, and different features impact on the classification results very different.

At present, short text classification research is still relatively less. And because of the explosive growth and popularity of micro-blog and other short textual content, many areas such as Internet content security[6], text retrieval, network-based marketing, Web mining and other information processing have required a short text feature selection and classification. Therefore, in order to facilitate users in various fields to use the short text data, the research on the feature extraction and classification has been very necessary.

II. RELATED WORK

Short text feature extraction methods and traditional long text feature extraction methods have in common. The key of ordinary text feature extraction is the accuracy of feature word, that is, the accuracy of segmentation and statistical. There are many corresponding algorithms for this such as N-Gram, methods based on vector space models and statistical and some improved algorithms [7]. However, because of the specific characteristics of short text we need to research and design more suitable methods.

Currently, there are methods based on combining statistical and rule (χ^2-Max and TF/IDF method), introduction of semantic paradigm to classify the short text, and intervened by artificial means. Le Wang proposed WR-KMeans method: add the related terms based on HowNet into each speech segment. Through this method extends the traditional TF/IDF model and, to a certain extent, alleviate the problems caused by excessively low Keyword Frequency [8].Hui He employs N-gram feature extraction to capture Chinese chunks from texts, which reflect the text semantic structure and character dependency. Then RPCL algorithm is applied to realizing text clustering, which doesn't need know the exact number of clusters [9]. However, this feature extraction method is based on N-Gram, the expansion of feature has a lot of limitations. YAN Rui proposed a dynamic combination classification algorithm for short text. Firstly, he constructs a tree structure of combined classifier, which can effectively alleviate the affect on the classification performance by the sparse and highly uneven short text feature [6]. HE Tao uses the Pinyin sequence extracted from

978-1-4244-5392-4/10/$26.00 ©2010 IEEE

N-gram fragment of Chinese words to form a network of short texts of Chinese characteristics. This approach takes N-gram fragment after Segmentation, and avoids the impaction on the clustering results by the deformation term [10]. YU Jinkai designed a gram correlation matrix to statistic and merge the feature words. Based on the fixed length N-Gram algorithm he can extract different length feature words [11].

III. SHORT TEXT FEATURE SELECTION

A. Feature Selection Model

In the traditional process of document classification, people separate certain document into a list of words and using some criterions to evaluate these words. It gives each word a score of its contribution to the classification. In the above process, we consider each word equally. However, this method ignores the importance of each word's part-of-speech. A sentence or a document makes up of several notional words and structrural words. No matter to the semantic meaning in a sentence or to relevance of document topic, noun and verbs hold much more information than preposition and other functional words. Those prepositions and other functional words only have the function of making the whole sentence fluent and complete.

Meanwhile, to a certain feature from a piece of micro-blog, its meaning is more concrete than the supervised category, hence we call these feature Concrete Feature. A supervised category is a more generalized concept, so we call these General Concept. There are many concrete features belonging to a certain general concept's set, as the following formula shows.

$$S(GeneralConcept_j) = \{ConceptFeature_1, \ldots, ConceptFeature_n\} \quad (1)$$

Hence, before our micro-blog classification, we will change those concrete features into its general concept, which not only reduces the whole number of feature space and avoids the sparse vector problem in each piece of micro-blog, but also improves the precision and efficiency. Based on the above analysis, we propose a feature selection model based on part-of-speech and HowNet in the area of micro-blog classification. See Fig. 1.

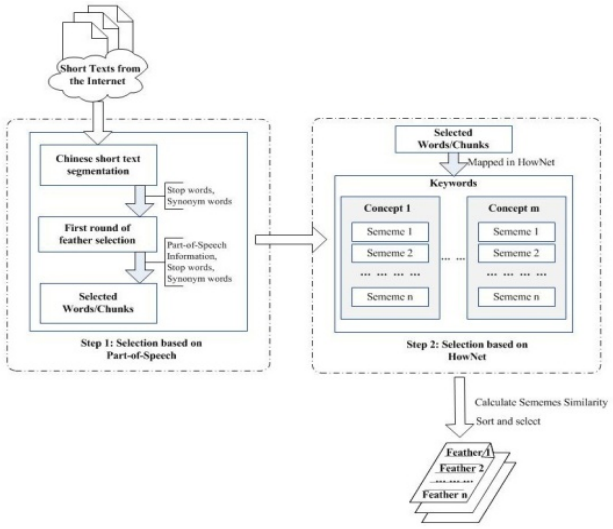

Figure 1. Feature selection model based on pos and HowNet

B. Selection based on Part-of-Speech

No matter to the semantic meaning in a sentence or to relevance of document topic, noun and verbs hold much more information than preposition and other functional words. However, in order to the sentence's fluency and integrity, a huge number of functional words used in one document. There are a large amount of functional words existing in the feature vectors. They not only cause longer time in classification, but also reduce the precision of classification. We do some statistical research in the public corpus of China Daily published in January, 1998. There are 1140931 Chinese terms after segmentation in this corpus. It contains 614451 functional words, which is 53.9% of the total terms.

To those segmented initial feature selection set, we use the following tags of part-of-speech in Table I to filter those initial features. We only reserve noun, verbs and adjectives.

$$\mathrm{Term_i} \equiv (\mathrm{Term^i_{LiteralVal}}, \mathrm{Term^i_{POS}})\ \mathrm{Term_{POS}} \in \{\mathrm{NOUN, VERB, ADJ}\} \quad (2)$$

C. Selection based on HowNet

1) What HowNet is

HowNet is a knowledge base uses the concepts which the Chinese and English words represented to reveal the relationship between concepts and attributes of the concept. It uses the selected Sememe and selected semantic relationships between concepts as the basic unit to describe the concept, and use a structured description language to describe a variety of concept precisely [12]. Based on the Biaxial Theory used to classify the concept of events, it fully reveals the complex relationships between things. Sememe is the basic, smallest and cannot be divided unit of meaning. We use the knowledge dictionary to define a DEF for each word entry. DEF terms defined by one or more Sememes, this is our conceptual access basis. Through the extraction of the defined Sememe of the words' DEF, we can get the short text feature and the rule of the short text. Then we will obtain more suitable concept as the features options of short text.

TABLE I. PART-OF-SEPPCH TAGS IN FEATURE SELECTION

General PoS Category	Specified PoS Tag	Explanation
ADJ	a	Adjective
	ag	Adjective Morpheme
	ad	Adverb
	an	Adnoun
NOUN	n	Noun
	nr	People's Name
	ns	Place's Name
	nt	Organization' Name
	nz	Other Proper Nouns
VERB	v	Verb
	vd	Avendo
	vn	Gerund

In HowNet, the concept and its description of each word format a record, and each record contains five main elements [13]: W_X = correspond concept of the word, G_X = part of speech of the word, S_E = emotion tendency of the word, E_X = example on the concept, DEF = concept definition (X = Chinese or English). Take the word which NO .= 016028 for example:

W_E = happy
G_E = ADJ
S_E = PlusFeeling |正面情感
E_E = N /A
DEF = {joyful |喜悦}

2) How to use HowNet

After pretreatment of the short text by Selection based on Part-of-Speech, we can obtain the following candidate feature set $S_{OriginalFeature} = \{Feature_1, \ldots, Feature_n\}$. For each $Feature_i$, which is isolated feature point, there is no semantic connection between them. If Feature_i belongs to the noun, verb or chunks defined in HowNet Wordlist, then we will expand the meaning of Feature_i at a concept and sememe level by HowNet, that is, we extend the original rare and isolate concepts in a semantic level and get a new set $S_{HowNetFeature}$ of relevant features supported by concept and sememe. If Feature_i is not in HowNet WordList, then we will add Feature_i into new feature set $S_{HowNetFeature}$ directly. Specific feature expansion method based on concept and sememe level of HowNet elaborates as follows:

STEP1: For each $Feature_i$ ($\text{Feature}_i \in S_{\text{HowNetWordList}}$), we get a set of concepts related to this feature and sememe sets of each concept corresponds by HowNet, namely:

$S_{Concept}(Feature_i) = \{Concept_1, \ldots, Concept_k\}$
$S_{Sememe}(Concept_j) = \{Sememe_1, \ldots, Sememe_m\}$

STEP2: For each element $Concept_j$ in $S_{Concept}(Feature_i)$, we calculate the semantic similarity w_{ij} between $Feature_i$ and $Concept_j$ by HowNet. We use L2 norm for each w_{ij}.

STEP3: Sorts the elements in the collection $S_{Concept}(Feature_i)$ according to semantic similarity w_{ij} between $Concept_j$ and $Feature_i$, take the former α Concepts as the most relevant Concept of Feature_i. if $|S_{Concept}(Feature_i)| < \alpha$, then in view of the semantic network structure of HowNet, we can use the obtained Concept as Feature, GOTO STEP1.

STEP4: For the α Concepts of $Feature_i$ obtained in STEP 3, we can get the correspond sememe sets S_{Sememe}, and for each element Sememe_t in the sememe sets, calculate the semantic similarity $dist_{Sememe_t}(Feature_i)$ between Sememe_t and $Feature_i$, which

$dist_{Sememe_t}(Feature_i) = w_{ij} + w_{ik}$ ($Sememe_t \in S^j_{Sememe}$ and $Sememe_t \in S^k_{Sememe}$)

STEP5: For each $Sememe_i$ we get by the above methods, we sort them by $dist_{Sememe_t}(Feature_i)$, then fetch the first β Sememes as the final feature elements of the nearest semantic expansion collection $S_{HowNetFeature}$ of Feature_i, and use $S_{HowNetFeature}$ to replace the original feature set $S_{OriginalFeature}$.

D. Basic Algorithm

In sum, to those short texts like micro-blog or short message, we first do word segmentation to those corpuses and filter useless words based on those part-of-speeches. Second, to a certain word we get from the first step, we use the HowNet to calculate and obtain the top α similar concepts. Third, we calculate these semantic similarities of those α sememes and the feature. Fourth, we get the top β nearest sememes from that feature and using these top β nearest sememes as the expended feature set to do short text classification.

Basic Algorithm is shown below:

```
ChineseWordSegmentation(DataSet)
PartOfSpeechFilter(DataSet)
For i = 0 to ShortTextNum
    For j = 0 to ShortText[i].FeatureNum
        GetRelatedConceptWithSimWeight
        (ShortText[i][j])
        SimWeightNormalization_Sort(α)
        GetNearestSememe(β)
DoTextClassification()
```

IV. EVALUATION

A. DataSet

We collect our test data from Sina micro-blog [14]. We take the advantage of the search function of Sina micro-blog and manually choose 400 short text covering four categories (Finance, Auto, Military, Sports). We use the K Nearest Neighbor classification algorithm to do our experiment.

B. Measurement

To balance the influence between precision and recall in traditional text classification, we will adopt the F_1- Measure as our precision's criterion. Meantime, considering that both the precision and recall are calculated according to certain one category and it only represents a local effect, we use Macro-Averaging and Micro-Averaging based on F_1 - Measure to evaluate the performance of our classifier in a global aspect.

To a certain known category K_i and the classifier decided category C_i, the precision, recall and their F_1 are calculated by the following formulas.

$$\text{Precision}(K_i, C_i) = \frac{|K_i \cap C_i|}{|C_i|} \quad (3)$$

$$\text{Recall}(K_i, C_i) = \frac{|K_i \cap C_i|}{|K_i|} \quad (4)$$

$$F_1(K_i, C_i) = \frac{2 * \text{Recall}(K_i,C_i) * \text{Precision}(K_i,C_i)}{\text{Precision}(K_i,C_i) + \text{Recall}(K_i,C_i)} \quad (5)$$

In the above formulas, $|C_i|$ denotes the number of documents which are classified into category C_i; $|K_i|$ denotes the number of documents whose category is K_i; $|K_i \cap C_i|$ denotes the number of documents which is classified into correct category.

We use Macro-Averaging and Micro-Averaging to evaluate each category's $F_1(K_i, C_i)$, see the following formulas.

$$\text{MacroAvg}_{F_1} = \frac{\sum_{i=1}^{|C|} F_{1i}}{|C|} \quad (6)$$

$$\text{MicroAvg}_{F_1} = \frac{\sum_{i=1}^{|C|}(\text{Recall}(K_i,C_i) * \text{Precision}(K_i,C_i))}{\sum_{i=1}^{|C|}(\text{Recall}(K_i,C_i) + \text{Precision}(K_i,C_i))} \quad (7)$$

Where |C| denotes the number of categories in the whole documents set.

C. Experiment Result And Analysis

The Macro-Averaging of F_1 based on several famous feature selection methods is depicted in Fig 2. From this figure, we can find that when we select 100 features, there is no significant difference in these four feature selection methods. They all stay in the same level. With the growth of number of features, the Macro-Averaging of F_1 of our feature selection methods based on HowNet is much higher than other three feature methods. When the feature number reaches 1000, which is so sparse to a piece of micro-blog, the values of Macro-Averaging of F_1 of four feature selection methods reduce to a relative low level.

In the performance of Micro-Averaging of F_1, when the feature space is between 200 and 400, the feature selection method based on document frequency has a better classification result that other two traditional methods. Our method based on HowNet has the highest Micro-Averaging of F_1 when feature number is from 200 to 1000, which demonstrates that our method always has the best performance. See Fig 3.

V. CONCLUSION

The main contribution of this paper is: First, considering the characteristics of Chinese short text, we propose our feature selection model in Chinese short text. Second, based on the syntactic attributes of Chinese language, we use part of speech to filter our candidate features. Meanwhile, we utilize the relevance of the semantic web from HowNet to expand the separate concrete concepts, which improve the classification precision in Chinese short text. In the future, experiments will explore the optimal combination of various thresholds in the algorithm to further enhance the short text classification quality and apply this approach into some web application.

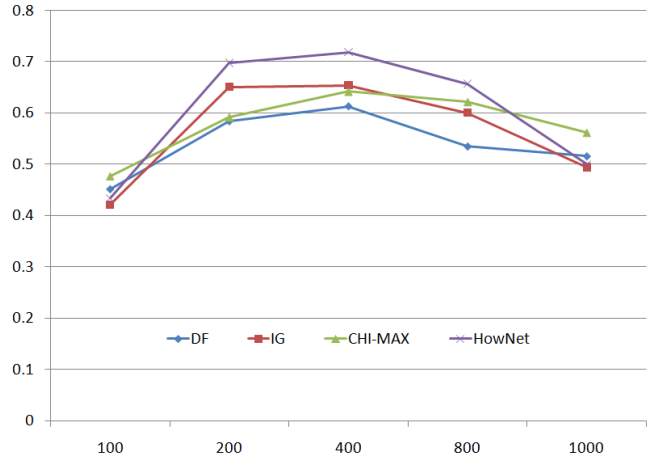

Figure 2. Macro-Averaging of F_1 comparision by four feature selection methods

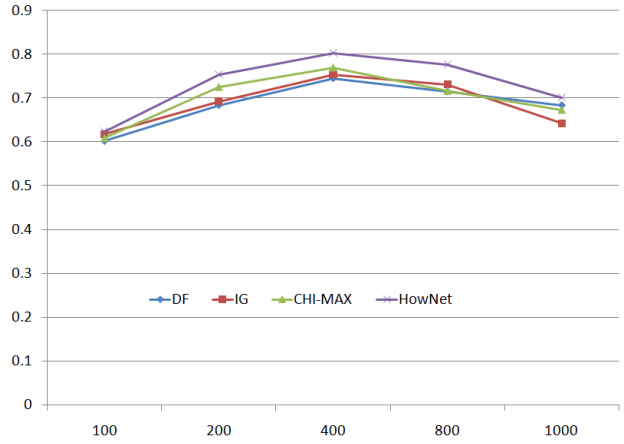

Figure 3. Micro-Averaging of F_1 comparision by four feature selection methods

REFERENCES

[1] Qindong Sun, Qian Wang, Hongli Qiao. The Algorithm of Short Message Hot Topic Detection Based on Feature. Information Technology Journal, 8(2): 236-240, 2009.

[2] Yang, Y. Expert network: effective and efficient learning from human decisions in text categorization and retrieval. In Proceedings of 17th ACM International Conference on Research and Development in Information Retrieval (SIGIR 1994), pp. 13–22.

[3] Grinev Maxim, Grineva Maria, Boldakov Alexander, Novak Leonid, Syssoev Andrey, Lizorkin Dmitry. Sifting micro-blogging stream for events of user interest.Proceedings - 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009), pp. 838.

[4] Jansen Bernard J., Zhang Mimi, Sobel Kate, Chowdury Abdur. Micro-blogging as online word of mouth branding. Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems (CHI 2009), pp.3859-3864.

[5] Shen Yang, Tian Chengeng, Li Shuchen, Liu Shichao. The grand information flows in micro-blog. Journal of Information and Computational Science, 6(2):683-690, 2009.

[6] Yan Rui, Cao Xian-Bin, Li Kai. Dynamic Assembly Classification Algorithm for Short Text. Tien Tzu Hsueh Pao/Acta Electronica Sinica, 37(5):1019-1024.

[7] Aleahmad Aholfazl, Hakimian Parsia, Mahdikhani Farzad, Oroumchian Farhad. N-gram and local context analysis for Persian text retrieval. In Proceeding of 9th International Symposium on Signal Processing and its Applications(ISSPA 2007),pp. 1-4.

[8] Wang Le, Jia Yan. Instant message clustering based on extended vector space model. 2nd International Symposium on Intelligence Computation and Applications (ISICA 2007), pp. 435-443.

[9] He Hui, Chen Bo, Xu Weiran. Short Text Feature Extraction and Clustering for Web Topic Mining. 3rd International Conference on Semantics, Knowledge, and Grid(SKG 2007), pp. 382-385.

[10] He Tao, Cao Xian-Bin, Tan Hui. An Immune Based Algorithm for Chinese Network Short Text Clustering. Zidonghua Xuebao/ Acta Automatica Sinica, 35(7):896-902, 2009.

[11] YU Jinkai, Wang Yingxue, Chen Huaichu. An Improved Text Feature Extraction Algorithm Based on N-Gram. Library And Information Service, 48(8):48-50, 2004.

[12] Dong Zhendong, Dong Qiang. Theoretical Findings of HowNet. Journal of Chinese Information Processing, 2007.7(4):36-43.

[13] HowNet: http://www.keenage.com

[14] Sina Micro-Blog: http://t.sina.com.cn