2008 International Conference on Computational Intelligence and Security

A New Feature Extraction Approach Based on Sentence Element Analysis

Chuangxin Yang
School of Computer Science and Engineering,
South China University of Technology
Guangzhou, 510641, China
Guangdong University of Business Studies
Guangzhou 510320, China
yang@gdcc.edu.cn

Hong Peng, Jiabing Wang
School of Computer Science and Engineering,
South China University of Technology
Guangzhou, 510641, China

Abstract

Considering that each sentence element of a sentence or clause plays an important role in describing case and (or) object in documents, a feature extraction method based on sentence element is proposed in this paper. The method can extract feature terms from documents effectively and weight them accurately. It first extracts sentence elements from dependency relationships, and then selects and weights the terms according to the sentence elements. Experimental results on a public dataset prove the feasibility of our approach and demonstrate its advantage to feature extraction method based on part of speech.

1. Introduction

Feature extraction plays an important role in text categorization and clustering. It is effective in reducing dimensionality and noise. The popular methods for text feature extraction include evaluation function, feature correlation and semantic understanding.

As for evaluation function, Yang *et al*[1], conducted a comparative study of supervised feature selection methods in statistical learning of text categorization. These methods include: document frequency (DF), information gains(IG), mutual information(MI), x2 statistic(CHI), and term strength(TS). IG and CHI are found to be most effective in their experiments [1, 5]. As for feature correlation, Galavotti[2] proposed a feature selection method in automated text categorization based on negative evidence correlation. Huang[3] provides a multi-type features co-selection, and Xu[4] provided a method of feature selection based on expectation maximization and cluster validity.

With the development of natural language processing (NLP), more and more semantic understanding technologies have been applied in text feature extraction area. Jin and Miao[5] provided an algorithm of extracting text character based on the model of context framework. Zhao[6] presented an approach of Chinese text representation based on semantic and statistic feature. HU[7] proposed an independent semantic feature extraction algorithm based on short text. Based on concept extraction with a shielded level, Liao[8] proposed a feature selection method in Chinese text classification.

All of the methods mentioned above can reduce dimensionality of feature space effectively. Most of them must filter noise by stop word list or its part of speech before feature extracting. But it is inconvenient to build a universal stop word list. Some commonly used methods, such as part of speech, would result in feature lost. In this paper, we propose a new feature extraction method based on sentence element analysis. The major advantages of our method are:

(1) Extract feature set conveniently and effectively without pretreatment, such as constructing stop word list or filtering noise according to part of speech, etc.

(2) Select feature based on sentence element, our method has good performance and can be used in most cases.

The rest of this paper is organized as follows. In section 2, we systematically study similarities and differences among methods in statistical on terms, weight and part of speech. In section 3, we present a Sentence Element Extraction algorithm (called SEE), and employ this algorithm in Candidate Feature Set Extraction (CFSE). In section 4, our method is

978-0-7695-3508-1/08 $25.00 © 2008 IEEE
DOI 10.1109/CIS.2008.37

IEEE computer society

applied in IG. Finally, we test our feature set in classification task using KNN classifier and analyze the experimental result in section 5.

2. Comparative analyse on several feature extraction methods

Different from Ref[1], we conduct a comparative study on the common ground and difference of terms, weights and part-of-speech, among five feature sets. The purpose of this comparison is to choose algorithm in section 4, and to compare experiment results in section 5.

The five feature sets were acquired from five feature selection methods. These methods are information gain (IG), mutual information (MI), cross entropy (CE), x2 statistic (CHI), weight of evidence (WE), and right half of IG (RIG). All the methods train with the same corpus, in 1000 dimension feature space, and employing ICTCLAS to segment text. The results of analytical comparison are showed in Table1 and Table2.

The upper triangle in Table 1 shows the percentage of the same feature terms between two different feature sets. The lower triangle shows the sum of different values of the same terms between different feature sets. It indicates that IG holds the largest percentage of the same feature terms compared with other criteria. This is the reasons why we employ IG to test our feature extraction approach.

Table 1. The common ground and difference of feature terms among five feature sets

	IG	CE	RH	WE	CHI
IG		98.3	84.2	76.9	84.0
CE	117.33		82.5	77.9	84.6
RH	1168.14	1285.48		63.6	71.3
WE	1661.25	1588.96	2636.82		74.6
CHI	1272.90	123.84	2197.36	1959.19	

Table 2. The part of speech distribution proportion

	IG	CE	RH	WE	CHI
n	35.78	36.07	33.30	39.42	37.46
v	27.83	28.50	28.19	25.85	28.16
p	5.88	5.04	6.12	3.63	5.99
c	4.29	4.41	3.90	5.72	4.45
a	4.18	4.09	4.50	4.28	4.29
q	3.98	4.09	4.38	3.45	3.18
m	4.30	3.88	4.39	2.28	2.15
d	2.95	2.83	4.38	2.59	2.52
ns	2.21	2.27	2.07	2.97	2.41
nd	2.05	2.11	2.13	2.45	2.21
u	1.14	1.17	1.15	1.46	1.54
ws	1.16	1.22	0.94	1.59	1.50
nt	1.04	1.02	1.13	0.85	1.07
r	0.90	0.92	1.14	0.39	0.66
j	0.73	0.75	0.58	0.98	0.71
b	0.57	0.58	0.73	0.65	0.64
ni	0.54	0.56	0.48	0.79	0.55
nz	0.19	0.19	0.15	0.24	0.22
i	0.11	0.11	0.07	0.15	0.13
nh	0.05	0.05	0.06	0.19	0.07
k	0.10	0.11	0.09	0.05	0.04
nl	0.03	0.03	0.10	0.02	0.03

From Table 2, we can see clearly that the part-of-speech distribution proportions are very similar. On the other hand, although nouns and verbs hold the greatest proportion, other part-of-speeches hold approximate 38 percent proportion. If only extract the nouns or verbs from text, the feature of other part-of-speech would be lost. It is an inefficient method to extract all of substantives in texts, because substantives occupy 97.4 percent in vocabulary [9].

3. Feature extraction method

A description on dependency relation and an introduction of the basic idea of our method are given in this section. Two algorithms, Sentence Element Extraction algorithm (called SEE) and Candidate Feature Set Extraction algorithm (called CFSE) are proposed according to the idea. SEE is employed to select the feature terms, and is used in CFSE which weights the feature and builds the candidate feature set.

3.1 Dependency relation

Parsing was first proposed by French linguist L. Tesniere in his paper *Element de Syntaxe Structural* in 1959[10]. Dependency syntax reveals the syntax structure of each part of the text language unit through parsing its dependency relations. The primary content of dependency grammar structure is semantic dependency relation, which is the binary relation of a word pairs in a sentence. In the word pairs of the binary relation, one is denoted as *head*, the other is denoted as *dependent*. Dependency relations reflect the dependent relation of the *head* and the *dependent* in semantic.

Parsing plays an important role in areas such as automatic ASK & ANSWER, machining translation, information indexing, information extraction, etc. At present, Chinese dependency relations can be divided into syntax and semantic. HOWNET is a dependency relation parser based on semantic, with 55 semantic dependency relations. LTP (Language Technology Platform) is syntax parser developed by the Language Lab of Harbin Institute of Technology. LTP was employed in this paper, and its 27 syntax dependency relations are showed in Table 3.

Table 3. Dependency relation in parser

Sign	Relation	Sign	Relation
ATT	attribute	DE	DE structure
QUN	quantity	DI	DI structure
COO	coordinate	DEI	DEI structure
APP	appositive	BA	BA structure
LAD	left adjunct	BEI	BEI structure
RAD	right adjunct	ADV	adverbial
SIM	similarity	VOB	verb-object
MT	mood-tense	SBV	subject-verb
IS	indep. structure	VV	verb-verb
CMP	complement	CNJ	conjunctive
POB	prep-obj	IC	indep. clause
HED	head	DC	dep. clause

3.2 Sentence Element Extraction

The basic idea of our method is to extract feature in syntax lay, because sentence element is effective to present the meaning of most documents. As a part of sentences or clause, each of sentence elements gives its individual contribution on describing case and (or) object in text. For example:

Subject: Denotes the doer of the action or what is described;

Predicate: One of the two main constituents of a sentence, modifying the subject and including the verb, objects, or phrases governed by the verb;

Object: A noun or substantive that receives or is affected by the action of a verb within a sentence.

And attribute used as a modifier for subject or object, adverbial and complement for predicate.

An algorithm, called SEE, is provided to extract sentence elements from dependency relations in Table 3. Most sentence elements of words can be gained directly from their dependency relationship in the sentence, e.g. to determine the word as subject from SBV relationship. And others can be determined by their parent nodes, such as *verb-verb* and *appositive.* The algorithm will be used in CFSE to get the sentence element of each term from parsing tree. It is shown in detailed as follows:

Algorithm of Sentence Element Extraction

```
INPUT: Current term in sentence term. The term's dependency relation derela, position i and its parent node position j.

OUTPOUT: term and its sentence element senelem.

BEGIN
1) if (derela ==SBV)
       then senelem of term[i] is subject and term[j] is predicate;
2) if (derela ==VOB)
       then senelem of term[i] is predicate and term[j] is object;
3) if (derela =={ATT,ADV or CMP})
       then senelem of term[i] is {attribute, adverbial or complement};
4) if (derela ==APP or VV)
       then the senelem of term[i] is determined by the derela of its parent node term[j]
  END;
```

3.3 Candidate Feature Set Extraction

The goal of the Candidate Feature Set Extraction is to extract feature terms from training corpora. For this purpose, the SEE algorithm was applied after every sentence was parsed. We assign a weight to each feature, which is determined by its sentence element of feature terms. The reason is that one word can play different role in different text. The weight was gained from statistical result of IG method. Consequently, the feature value of each term depends on two aspects: the sentence element and the frequency of its appearance in training text.

The detailed algorithm is shown as follows:

Algorithm Candidate Feature Set Extraction

```
INPUT: Training documents; P[6][2]={(subject, $\lambda_1$), (predicate, $\lambda_2$) , (object, $\lambda_3$) ,
       (attribute, $\lambda_4$) , (adverbial, $\lambda_5$) ,
       (complement, $\lambda_6$)};
```

```
OUPUT: Candidate Feature Set CFS ;  //every
       feature term in CFS be denoted as (id,
       termset[],value).
BEGIN
1) int termcount, sencount;  // termcount denotes
       the total number of terms in a sentence;
       sencount is the amount of sentence in a
       document.
2) Paser();  // parsing training document;
3) for(int i=0;i< sencount; i++)
   {
4)   for(int j=0; j<termcount; j++)
     {
5)     Getword(j); GetParse(pair<int, char *>,
       wordIdx);
6)     if (pair.second∈P)
       then employ SEE to extract the sentence
       element of the jth term.
7)     if (the term found in CFSd_i )
       then add λ to its value, else create a new
       feature in the CFSdi. //λ's value contain in P.
     }
   }
   END;
```

4. IG algorithm based on sentence element

4.1 Algorithm Information Gain (IG)

Information gain is widely employed as a term-goodness criterion in the field of machine learning. It measures the number of bits of information obtained for category prediction by knowing the presence or absence of a term in a document. Let $\{d_1,d_2,\ldots,d_m\}$ denote the set of categories in the target space. The information gain of term t id defined as:

$$IG_{t_k}=H(D)-H(D|t_k) \tag{1}$$

$$H(D)=-\sum_{d_i\in D}^{n}(P(d_i)\times\log_2(P(d_i))) \tag{2}$$

$$H(D|t_k)=-\sum_{d_i\in D}^{n}(P(d_i|t_k)\times\log_2(P(d_i|t_k))) \tag{3}$$

Given a training corpus, the information gain of each unique term was computed in the training corpus. And some terms would be removed form the feature space because its entropy is less than some predetermined threshold.

In formula (2) and (3), there are several different methods to calculate $P(d_i)$, when IG was employed. Most of these methods are based on frequency of terms appearance. In Ref [10], P was set as follows:

$$P(d_i)=\frac{|wordset(d_i)|}{\sum_i|wordset[d_i]|} \tag{4}$$

The $wordset(d_i)$ means the amount of terms in the d_ith catalog. In Ref[11,12], $P(d_i)$ denotes the probability of documents appearance. In the methods introduced above, feature value of the terms was gained from its frequency of appearance in the d_ith catalog and absence in other catalogs.

4.2 IG Algorithm based on Sentence Element

In this algorithm, the candidate feature set in section 3 was applied in IG. The feature value of a term depends on its weight in candidate feature set and entropy. Therefore, the $P(d_i)$ and $P(d_i|t_k)$ in formula (2) and formula (3) will be modified.

Let $CFSd_i$ be the candidate feature set of the ith class, $CFSd_i.[t_k]$.value denote the weight of the kth feature t_k in class d_i, $CFSD$ is the candidate set of all class. $CFSD=\cup CFSDd_i$ (i=1,2,…,n, n is the total of training documents); $CFSD.[t_k]$.value denotes the weight of feature t_k in $CFSD$, that is

$$CFSD.[t_k].value=\sum CFSd_i.[t_k].value$$

(i=1,2,…,n, n is the number of classes):

$$P(d_i)=CFSd_i.value/CFSD.value \tag{5}$$

$$P(d_i/t_k)=CFSd_i[t_x].vlaue/CFSD[t_x].value \tag{6}$$

Substitute(5)and(6) into formula (1)(2)(3)，we can get our improved IG expression:

$$IG_{t_k}=-\sum_{d_i\in D}^{n}((CFSd_i.value/CFSD.value)\times\log_2(CFSd_i.value/CFSD.value))+$$
$$\sum_{d_i\in D}^{n}((CFSd_i[t_x].vlaue/CFSD[t_x].value)\times\log_2(CFSd_i[t_x].vlaue/CFSD[t_x].value)) \tag{7}$$

5. Experiment results

5.1 Experimental Design

We use a public corpus in the experiment to demonstrate the advantage of the proposed method. The corpus is composed of 1882 training documents and 935 testing documents. All the documents are divided into ten categories, including: environment, computer, communication, education, economy, military, sport, medicine, art and politics.

In order to prove the effectiveness of our methods, several approaches are used in this experiment to compare the experiment result. BSES is a method proposed in this paper that based on sentence element analysis, IG is a traditional Information Gains method , N&V and N is a improved IG based on part of speech filtering , the differences is that N&V chooses not only noun but also verb while N only chooses noun.

In this experiment, TF.IDF algorithm was employed by each method when weighting terms. We perform SEE and CFSE, extract candidate feature terms, and adopt TF.IDF IG to weight feature value of candidate features. We compare our method with TF.IDF IG and feature selection based on part of speech. We then employ KNN classifier to examine the validity of feature selection method.

5.2 Comparison and Analysis

Table 4 shows the average precision and recall of different methods. In precision, BSES is 1.2% better than IG, 2%better than N&G and N; In recall, the result is significant: BSES is 1% better than IG, but it is almost 4% far better than N&G and N. Fig.1 shows the precision curves of different methods according to each category, and Fig.2 shows the recall of them.

From the experiments, we can see that the precision and recall of BSES are both better than that of other methods. Compared with IG, we find that insignificant symbols such as ‘「’ , ‘」’ , ‘【’ don’t appear in the extracted feature set of BSES, besides the performance of BSES is better than that of IG. Compared with N&V and N, the performance of our method is much better than that of them.

Table 4. Average precision and recall of different methods.

	BSES	IG	N&V	N
precision	88.594	87.370	85.585	85.115
recall	86.532	85.645	82.927	82.697

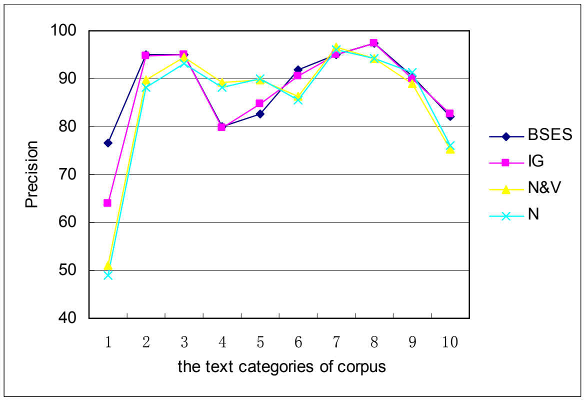

Figure 1. Precision curves of different methods

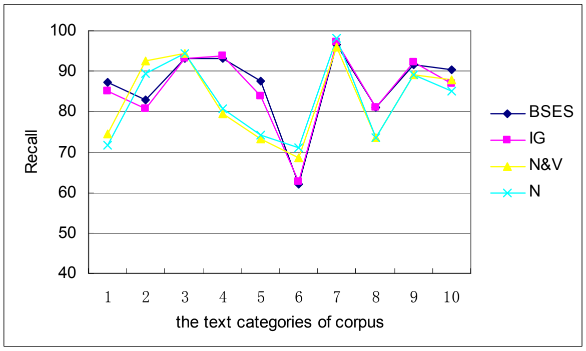

Figure 2. Recall curves of different methods

6. Conclusion

In this paper, we proposed a new feature extraction method based on sentence element analysis from syntactic perspectives. We demonstrated the effect of the method using IG and KNN classifier. Extensive experiments conducted on a public corpus show that our method is better than the others method. Besides, the method can be optimized in several aspects, e.g. using anaphora resolution to make sentence element definite and adjusting λ in CFSE to gain the better result.

Our future effort is to apply our method to other classifiers, and seek new approach which combines syntactic and semantic in text feature extraction.

Acknowledgment

This work was supported by the Guangdong Natural Science Foundation under Grant 06300170. We would like to thank Jia Wei, Guiduo Yan, Qinyang

Li and Wenliang Xie for conducting the experiment and fruitful discussions. We also like to thank HIT-IR-Lab for providing Language Technology Platform.

References

[1] Yang Y, Pedersen J. A Comparative Study on Feature Selectionin Text Categorization[C]//Proceedings of the 4th International Conference on Machine Learning. Nashville: Morgan Kaufmann Press, 1997:412-420.

[2] Galavotti L, Sebastiani F, Simi M. Feature Selection and Negative Evidence in Automated Text Categorization[C]//Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining. Boston: ACM Press, 2000.

[3] Huang S, Chen Z, Yu Y, et al. Multitype Features Coselection for Web Document Clustering[J]. IEEE Trans Knowledge and Data Engineering, 2006, 18(4):448-459.

[4] XU Junling, XU Baowen, et al. A New Feature Selection Method for Text Clustering. Wuhan University Journal of Natural Sciences Vol.12 No.5 2007 912-916.

[5] JIN Yao-Hong, MIAO Chuan-Jiang, An Algorithm of Extracting Text Character Based on a Model of Context Framework, Journal of Computer Research and Development, Apr1 2004, Vol141 ,No14.

[6] ZHAO Peng, GENG Huan-tong, An Approach of Chinese Text Representation Based on Semantic and Statistic Feature, Journal of Chinese Computer Systems, Vo l128 No. 7 2007.

[7] HU Jia-ni, GUO Jun, et al. Independent Semantic Feature Extraction Algorithm based on Short Text, Journal on Communications, December 2007, Vol.28 No.12.

[8] LIAO Sha-sha, JIANG Ming-hu. A Feature Selection Method in Chinese Text Classification Based on Concept Extraction with a Shielded Level Journal of Chinese Information Processing, Vol. 20 ,No. 3. 2006.

[9] YU Shi wen, ZHU Xue feng, et al. the Grammatical Knowledge base of Contemporary Chinese. Institute of Computational Linguistics; Peking University. 1997.

[10] L. Tesniere. Element de Syntaxe St ructural [M] .Paris : Klincksieck , 1959.

[11] LU Song LI Xiao-li et al. An Improved Approach to Weighting Terms in Text, Journal of Chinese Information Processing, Vol. 14 ,No. 6. 2001.

[12] Songbo Tan et al. A Novel Refinement Approach for Text Categorization. ACM CIKM 2005.