Geo-spatial Information
Science (Quarterly)

Volume 6, Issue 4, Page 66-72
December 2003

An Effective Concept Extraction Method for Improving Text Classification Performance

ZHANG Yuntao GONG Ling WANG Yongcheng YIN Zhonghang

KEY WORDS text classification; concept extraction; characteristic term; association rule; algorithm

ABSTRACT This paper presents a new way to extract concept that can be used to improve text classification performance (precision and recall). The computational measure will be divided into two layers. The bottom layer called document layer is concerned with extracting the concepts of particular document and the upper layer called category layer is with finding the description and subject concepts of particular category. The relevant implementation algorithm that dramatically decreases the search space is discussed in detail. The experiment based on real-world data collected from InfoBank shows that the approach is superior to the traditional ones.

1 Introduction

With the widespread and increasing availability of text documents in electronic form, it becomes more critical to use automatic methods to analyze, organize and manage text documents. The huge scale and unstructured nature of electronic text cause an information overload for users. Therefore, automatic text classification is becoming a hot research field. The automatic text classification based on their contents refers to the task of automatically assignments text documents (hereafter shortly document) into one or more predefined categories. It can be considered as the form of text mining in the sense that it abstracts the key contents of free-text documents into a single category label[1]. The automatic classification of documents will help users to prioritize the relevance of documents to the user's needs. The concept-based approach is the hotspot in the field of automatic text classification. However, the topic concept extraction from text or particular category is a difficult task because the natural language is ambiguous and inaccurate.

2 Concept representation of text and category

The text can be treated as "bag of words" or "set of words". However, many words in text are non-informative. A document contains one or several conceptual topics (or called topic concept, hereafter shortly concept). It can be seen that concept can be represented by a group (combination) of particular meaning items of some

Project supported by the National Natural Science Foundation of China (No. 60082003) and the National High Technology Research and Development Program of China (No. 863-306-ZD03-04-1).

ZHANG Yuntao, lecturer, Ph. D candidate, Network & Information Center, Shanghai Jiaotong University, Shanghai 200030, China.
E-mail: ytzhang@mail.sjtu.edu.cn; yuntao_zhang@hotmail.com

characteristic terms. (Hereafter, the characteristic term implies the particular concept, rather than a common word or a phrase). The document can be identified by a set of characteristic terms. Furthermore, there are synonymous, near synonymous, polysemous or associative relationships among these characteristic terms representing a particular concept. It is said that one particular concept can be relevant to several characteristic terms. Moreover, a multiplex concept consists of simple concepts. A concept expresses more concrete and accurate meanings than keywords do. Fewer features are required to represent a text document as a set of concepts, rather than as a set of words.

Similarly, category can be identified by a set of concepts (or corresponding groups of characteristic terms) as well. Text classification usually focuses on topic concepts. Obviously, the set of characteristic terms identifying category is not simply conjunction of a series of sets of characteristic terms identifying document within particular category. The documents will be classified into category that contains the same or similar topics of document. We consider it as text classification based on concepts.

Concept extraction is the key technique of text classification based on concepts. In this research, the concepts defined as a set of measurable patterns are used to discover the relationship between particular document and particular class. A set of concepts is extracted from the input documents (training corpus).

In the approach presented in this paper, the computational measure is considered as two layers. The bottom layer (called document layer) is concerned with extracting the concepts of particular document and the upper layer (called category layer) is concerned with finding the description and subject concepts of particular category.

The document layer extracts the characteristic terms representing concept. In other words, the step will delete the non-informative terms from the feature space of document. The objectives of document layer are twofold. One is to prevent the over-fitting. Classifiers which over-fit the training data tend to be extremely good at classifying the data they has been trained on, but are remarkably worse in classifying other data[2]. Another is to reduce the cost of calculation when the unseen test document is classified into the predictive category. The characteristic terms represent the association relationship between particular group of characteristic terms (or called characteristic term chain) and particular category. The category layer is used to find the combination relationship among the characteristic terms.

3 Concept extraction in document

In document, the high-frequency words are more important to represent the content than low-frequency words. However, some high-frequency words have low content discriminating power such as prepositions, conjunctions, articles and pronouns. These high-frequency words are non-contextual words occurring in the text. They do not directly contribute to the content. Therefore, they are listed in the stop-list. It is clear that those words appearing in the stop-list will be deleted from the set of characteristic words before concept extraction. The preprocessing step is necessary for Chinese as well as English.

For Chinese, another preprocessing step is to change all of single-byte characters into double-byte ones. However, another step of preprocessing English document is stemming. By stemming, each word will be regarded as word-stem form in document. For example, "development", "developed" and "developing" will be all treated as "develop". Similarly to the stop-list, stemming reduces the amount of dimensions and enhances the relevancy between word and document or category. Therefore, the described method sums up

the number of times that term appears as itself or its morphological derivations when we calculate the weight of characteristic term representing topic concept within a particular document.

There are several important relationships among concepts represented by characteristic terms. For compromising the advantage and disadvantage of concept cluster, the synonymous terms and near-synonymous terms are preprocessed. All of them will be considered as a term when the concept is extracted from document.

The individual characteristic term used to represent topic concept is informative. The weight of characteristic term representing topic concept within a particular document is based on the relevancy between term and the topic concept of document which the term occurs in. It is proposed to extract characteristic term according to its contribution to topic concept of document. The chosen characteristic terms within a document have a contribution that exceeds the minimum contribution threshold. The threshold can be preset and adjusted.

We introduce the average frequency information, position information and length of terms in weighting equation. The equation is

$$W(t_i,d_j) = (N(t_i,d_j)+A)\log_2\frac{N(corpus)}{N(d \mid t_i)} \left(0.4+0.6\times\frac{L(t_i)}{\mathrm{MAX}(d_j)}\right)\frac{\frac{N(t_f,d_j)}{N(all,d_j)}}{\frac{N(t_f,corpus)}{N(all,corpus)}} \quad (1)$$

where $W(t_i,d_j)$ denotes the weight of term t_i in text d_j; $N(t_i,d_j)$, the number term t_i in text d_j; $N(corpus)$, the total number of texts in the corpus; $N(d \mid t_i)$, the number of texts containing term t_i in the corpus; $L(t_i)$, the length of term t_i; MAX (d_j), the maximum length of term in text d_j; $N(all, d_j)$, the total number of terms in text d_j; $N(t_i,corpus)$, the number of the term t_i in the corpus; $N(all, corpus)$, the total number of terms in the corpus and A denotes the adjustment value of term t_i according its position in text d_j. We estimate the adjustment value by experiments. The correlation between adjustment value and its position in text is listed in Table 1 where n denotes the number of term t_i in corresponding position. The adjustment value A of term t_i equals to the sum of adjustment values in all positions.

Table 1 Correlation between position and adjustment value

Term position	Adjustment value
Title	$n\times4$
Abstract	$n\times2$
Subtitle	$n\times1$
Reference	$n\times1$
Other	0

The weights of all terms in text are calculated by Eq. (1). The terms whose weight are greater than threshold are chosen as characteristic terms representing topic concept in the text. It is observed that a text document usually contains only one or several topic concepts that can be represented by a small number of characteristic terms. It will be used to explain the surprising phenomenon appearing in Lewis's experiment[3] where the small number of features is optimal and classification effectiveness drops off when a very large feature set is used. The words associating weakly with top concepts are harmful for the text classification when it is weighted at the feature computation. By experiment, we have found that the suitable number of characteristic terms extracted is between 15 and 20.

4 Concept extraction among categories

The most important step for the text classificat is to extract concepts within a particular category. Each text is viewed as a transaction, while a set of characteristic terms in the text will be considered as a set of items in the transaction[4]. The category of document is considered as attribute of document. The category of document is processed as item in the presented mining association algorithm. Consequently, the mining of multi-dimensional as-

sociation rules is transformed into the mining of single-dimensional association rule.

When the number of the characteristic terms within a particular category is G_k, the number of all combinations of characteristic terms chains is $2^{G_k}-1$. It will cost a great expense and is infeasible to investigate all the combinations of characteristic terms. Therefore, we design a novel algorithm to reduce the search space.

The presented algorithm is different from Apriori algorithm which is an influential algorithm for mining frequent itemsets for Boolean association rules. However, we think that the anti-monotone used in Apriori algorithm is benefitial to solve the problem. Anti-monotone means that all of its supersets will fail to pass the same test as well when a set cannot pass a test[5]. We use variation of anti-monotone property to prune the search space. In other words, if a set pass a test, all of its subsets will pass the same test as well.

The following procedure shows the pseudocode for the algorithm extracting concepts by the candidate groups of characteristic terms obtained from the texts of a particular category.

Algorithm: concept extraction based on candidate groups of characteristic terms for designated category

Input: formatted transactions set containing characteristic terms and category of document (computed by Eq. 1), minimum confidence threshold T_d

Output: groups of characteristic terms representing different concepts

Method:

1) for ($k = 1$; $k = C_N$; k++) { /* C_k is the category; k is between 1- C_N, C_N is the total number of different categories */
2) $C_{ck} = \phi$; // C_{ck} is the set of groups of characteristic terms for category C_k
3) G_k =obtain_set_of_distinct_characteristic_term_group_and_count (C_k) ;
4) C_a =obtain_set_of_all_characteristic_term(C_k);
5) for ($p = 1$; $p = M$; p++) { /* M is the maximum length of group characteristic term among G_k */
6) $D_{pk} = \phi$; /* D_{pk} is the set of groups of characteristic terms containing p characteristic terms for category C_k */
7) C_{ck} =obtain_p-characteristic-term (C_a, p, G_k) ;
8) }
9) $C_{ck} = C_{ck} \cup G_k$; /* the reminder groups of characteristic terms of G_k is confidence groups of characteristic terms according to the definition of group of characteristic terms because the documents with same group of characteristic terms will be classified into same category. */
10) }
11) return all C_{ck}, $k \in 1 \sim C_N$;
procedure obtain_p-characteristic-term (C_a, p, G_k) ;
12) if $p = 1$ then
13) for each $c \in C_a$
14) if confidence_confirm(c, C_k, T_d) then {
15) $C_{ck} = C_{ck} \cup c$; // add c into set of characteristic terms for category C_k
16) $C_a = C_a - c$; // prune step
17) G_d =subset (G_k, c) ; /* finding groups of characteristic terms containing characteristic term c among G_k */
18) $G_k = G_k - G_d$; // prune step
19) else $D_{pk} = D_{pk} \cup c$;
20) else for each $d_1 \in D_{p-1,k}$
21) for each $d_2 \in D_{p-1,k}$
22) if ($d_1[1] = d_2[1]$) $\wedge$ ($d_1[2] = d_2[2]$) $\wedge$ ··· $\wedge$ ($d_1[p-1] = d_2[p-1]$) then {
23) $q = d_1$ join d_2 ; // generate candidate group of characteristic terms
24) G_d =subset (G_k, q) ; /* finding groups of characteristic terms containing q */
25) if $G_d <> \Phi$ then
26) if confidence_confirm (q) then {
27) $C_{ck} = C_{ck} \cup q$; // add q into set of characteristic terms for category C_k
28) $G_k = G_k - G_d$; // prune step
29) }
30) else $D_{pk} = D_{pk} \cup q$;

31) }
32) return to C_{ck} ;
33) procedure obtain_set_of_all_characteristic_term(C_k)
34) $G_k = \phi$; // G_k is the set of set of characteristic terms representing concept of a text
35) for $\forall\ d \notin G_k$
36) $G_k = G_k \cup d$
procedure confidence_confirm (c, C_k, T_d)
37) confidence (c,C_k) = number_of_documents_containing_c_among_C_k/number_of_documents_containing_c_among_corpus
38) if confidence(c, C_k) $>$ T_d then
39) return to true
40) else return to false

The main idea is described as follows.

① The sub-concepts appearing in the text may not be the topic concept of the corresponding category, while the topic concepts of the category appear in some texts without failure.

② Some concepts appearing in the texts are not the essential concepts of the particular category, but the combination of them is.

The main steps in the algorithm are:

① Obtain the set of all characteristic terms representing concepts appearing in the texts of the particular category (referring to 3) in algorithm).

② Obtain the set of the characteristic terms representing topic concepts in the text category independently (referring to 4) and 33)).

③ Obtain the individual characteristic terms representing concepts independently (referring to 12)) and the combination of characteristic terms (referring to 20)). The key is to prune the different permutations of characteristic terms (referring to 16)) and to generate the candidate combination of characteristic terms (referring to 22)).

Based on the algorithm, the inner relationship of the individual documents will emerge and the subjects of categories be found. The word description of a concept may vary, but the characteristics of concept are fixed[5]. Characteristic terms can be used to represent the topics.

The confidence and support of characteristic terms are the dominant measure for text classification. Each measure can be associated with a threshold that can be adjusted by user aiming at different types of documents corpus. We specify the dominant measure as follows:

$$\begin{aligned}\mathrm{dom}(w_j, c_m) &= f(\mathrm{conf}(w_j, c_m),\ \sup(w_j)) \\ &= \begin{cases} 1 & ((\mathrm{conf}(w_j, c_m) \geqslant \mathrm{threshold}) \wedge (\sup(w_j) \geqslant \mathrm{threshold})) \\ 0 & ((\mathrm{conf}(w_j, c_m) < \mathrm{threshold}) \wedge (\sup(w_j) < \mathrm{threshold})) \end{cases}\end{aligned} \tag{2}$$

where conf(w_j, c_m) is the confidence of characteristic word w_j on peculiar category c_m; sup(w_j) is the support of characteristic word w_j .

5 Experiments and performance evaluation

The precision on the particular category c_m is the percentage of the number of correctly classified documents among category c_m divided by the total number of documents among category c_m :

$$\mathrm{precision}(c_m) = \frac{\mathrm{Assign}(c_m, c_m)}{N(\mathrm{all}, c_m)} \tag{3}$$

where Assign(c_m, c_m) represents the number of correctly classified documents among category c_m; $N(all, c_m)$ represents the number of documents classified into category c_m .

Recall on the particular category c_m is the percentage Assign (c_m, c_m) divided by $S(c_m)$:

$$\mathrm{recall}(c_m) = \frac{\mathrm{Assign}(c_m, c_m)}{S(c_m)} \tag{4}$$

where $S(c_m)$ represents the total number of documents that should be within the category c_m.

The experiment is based on real-world data collected from InfoBank, the biggest Chinese information base. The taxonomy in the base includes three layers and 196 categories. In the experiment, we compare the performance of classification using the concept extraction with that without using the concept extraction.

Experts can decide which documents are relevant to each topic and we assume that the

experts' judgments are correct. The sample set is divided into training and testing set. By selecting the 186 out of total of 196 categories that contain at least one training and one testing texts, there are 129 903 training texts and 6 152 testing documents.

In the experiment the most commonly used performance measure, namely precision, recall and F_1 are adopted. Here F_1 is a harmonic average of precision and recall:

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (5)$$

We compute the macro-average and micro-average to evaluate the performance across categories. The macro-average precision (similarly for recall and F_1) is obtained from the calculation of the precision for each category, and then their average. The micro-average is obtained by calculating the precision and recall of all categories. The macro average gives equal weight to categories, while the micro average gives the equal weight to the individual texts[6]. The precision, recall and F_1 in the open test (unseen texts) are shown in Table 2.

Table 2 Performance comparison/%

	Macro-average		Micro-average	
	Without concept extraction	With concept extraction	Without concept extraction	With concept extraction
Recall	89.73	92.52	90.15	93.69
Precision	81.22	85.76	81.14	86.27
F_1	85.26	89.01	85.41	89.82

From the comparative results, the classification approach using the concept extraction improves both precision and recall. Unfortunately, it is observed that the performance of several categories is worse, for example, the detailed micro average recall shown in Fig. 1.

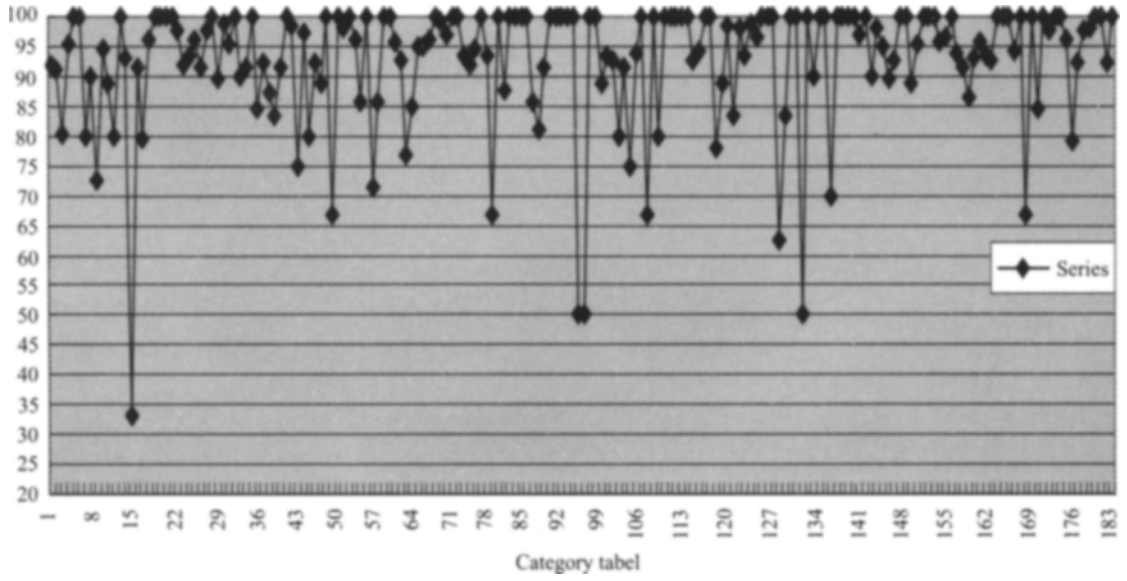

Fig. 1 Detailed micro-average recall

After the sample data are analyzed, we have found that the categories with worse performance usually contain less training set. We believe that the performance will be further improved for extended magnitude of the training text.

6 Conclusions

The characteristic term chains representing topic concepts of document or category can be generated in the way. The conceptual term chains have semantic significance for text classification.

The concept extraction for text classification improves the classification accuracy and produces a good prediction performance for test corpus because the presented approach decreases the dimensionality of the feature

space for the text classification. The result of experiment demonstrates the validity of this approach and corresponding algorithm. Moreover, this concept extraction approach is language independent and domain independent for text classification.

References

1 Tan A H (2001) Predictive self-organizing networks for text categorization. The 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Hong Kong.

2 Sebastiani F (2003) Machine learning in automated text categorization. ACM Computing Surveys. http://www. cvc. uab. es/shared/teach/a20368/ACMCS00. pdf.

3 Lewis D D (1992) Feature selection and feature extraction for text categorization. Speech and Natural Language Workshop, San Francsico.

4 Han J W, Kamber M (2001) Data mining: concepts and techniques. California: Morgan Kaufmann.

5 Li C, Luo Z S, Li Y H (2002) Research on automatic classification of documents based on concept attributes. 2002 IEEE International Conference on Systems, Man and Cybernetics.

6 Bakus J, Kamel M, Carey T (2002) Extraction of text phrases using hierarchical grammar. The Fifteenth Canadian Conference on Artificial Intelligence (AI'2002),Ottawa.

(Continue from Page 38)

References

1 Hou P(2001) Contours' vectorization in color topologic map Image by considering topology relationship: [Ph. D dissertation]. Wuhan: Wuhan University. (in Chinese)

2 Chen X Y, Li D R (1991) Automatic transformation of CCD scanned contours from raster to vector. *Acta Geodaetica et Cartographica Sinica*, 20:15-16(in Chinese)

3 Ma F(1995) A CCD scanned contour map recognition system on Windows. *Journal of Wuhan Technical University of Surveying and Mapping*, 20(3): 228-233 (in Chinese)

4 Mei X L, Zhang Z X, Zhang J Q (1995) Global broken contour connection through maximal clique graph search based on relational structural constraints. *Journal of Wuhan Technical University of Surveying and Mapping*, 20:101-105(in Chinese)

5 Cronin T(1995) Automated reasoning with contour maps. *Computers & GeoSciences*, 21:609-618

6 Chen R(2001) Identification of spot elevation in scanned map and automatic labeling contour lines: [Ph. D dissertation]. Wuhan: Wuhan University. (in Chinese)

7 Hao X Y(2001) Map information recognition and extraction technique. Beijing: Publishing House of Surveying and Mapping. (in Chinese)

8 Yan L(1999) Research on graphical elements of topographical map recognition techniques: [Ph. D dissertation]. Wuhan: Wuhan Technical University of Surveying and Mapping. (in Chinese)

9 Huang W J(2000) Theory and method of map pattern recognition. Beijing: Publishing House of Surveying and Mapping.

10 Ge Y H(2001) Research on techniques of topographical map recognition and map drawing: [Ph. D dissertation]. Wuhan: Wuhan University. (in Chinese)