Research Journal of Applied Sciences, Engineering and Technology 5(5): 1539-1543, 2013
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: June 28, 2012 Accepted: August 08, 2012 Published: February 11, 2013

A Text Categorization Algorithm Based on Sense Group

Jing Wan, Huiming Luo, Junkai Yi and Yubin Zhou
College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China

Abstract: Giving further consideration on linguistic feature, this study proposes an algorithm of Chinese text categorization based on sense group. The algorithm extracts sense group by analyzing syntactic and semantic properties of Chinese texts and builds the category sense group library. SVM is used for the experiment of text categorization. The experimental results show that the precision and recall of the new algorithm based on sense group is better than that of traditional algorithms.

Keywords: Semantic concept, sense group, support vector machine, text categorization

INTRODUCTION

Text categorization is an automatic processing that assigns a free text to one or more predefined classes or categories based on its content. Compared with English text categorization, the study of Chinese text categorization started later. And Chinese text categorization mostly makes use of the algorithms of English text categorization. As long as the internal structure of language is concerned, English is a hypotactic language, whereas Chinese is a paratactic language. But the current approaches for Chinese text categorization do not involve syntactic and semantic analysis and often make extraction and matching on the word level, with low categorization accuracy (Dai *et al.*, 2004).

Sense group, in the narrow sense, refers to a meaningful unit of words that are structurally and semantically related to each other. In the wider sense, sense group is the combination of associated concepts. More accurate identification of sense group means easier subject identification (Zhou *et al.*, 2004).

In order to overcome defects of the existing algorithms of Chinese text categorization, this study, considering the features of Chinese language and using the semantic dependency parsing put forward in reference (Li *et al.*, 2004), proposes a text categorization algorithm based on sense group. Sense-group-based text categorization algorithm trains the corpus according to syntactic and semantic features and builds a category sense-group library. In light of the categorization with sense group as the unit, these sense groups of the text to be categorized are extracted. Then these categorization attributes of the text are obtained using Support Vector Machine (SVM). As the sense-group extraction considers the syntactic and semantic features of the Chinese language, the text is presented in a way that is more adapted to the human's mode of thinking. Thus, the meaning of the text is better grasped by a computer, which can performs text categorization with higher precision.

SENSE GROUP AND TEXT CATEGORIZATION ALGORITHM

Sense group is a meaningful unit of words that are structurally and semantically related to each other. Sense group, combining a group of concepts together according to certain association, represents the intended meaning through a cluster of such interrelated units of words (Chen and Martha, 2004). How to acquire the concept, represent these sense groups and build the category sense-group library are the key processes in the algorithm for sense-group-based text categorization. The flowchart of the establishment of category sense-group library is shown in Fig. 1.

First, ICTCLAS30 (Zhang and Liu, 2002) is used to process the training texts in the corpus for word segmentation and part-of-speech tagging. The obtained results are then subjected to syntactic analysis. Based on the rules of Chinese syntactic understanding, the weights of the words in the clauses are assigned depending on the importance of the clauses to the text. Suppose the training text is T_i and the text after syntactic analysis is:

$$T_i=\{(d_1, w_1),(d_2, w_2), \ldots ,(d_n, w_n)\}$$

where,

Corresponding Author: Junkai Yi, College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China

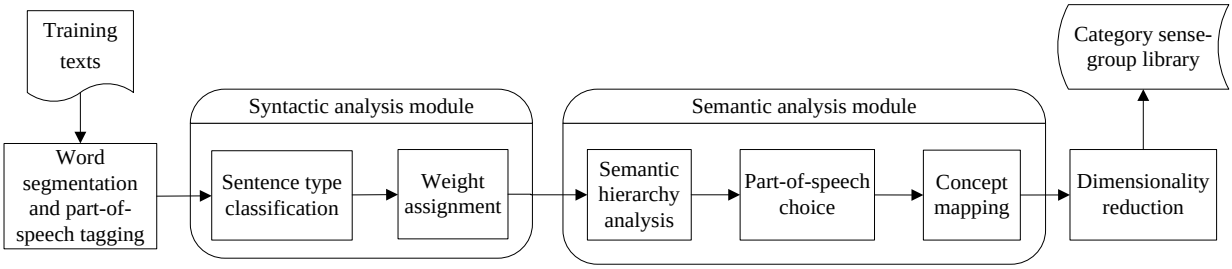

Fig. 1: Flowchart of the establishment of category sense-group library

d_i : A word
w_i : The weight of word

Semantic hierarchy analysis is performed on the results of syntactic analysis with assigned weights. Considering semantic dependency in Chinese sentences (Li *et al.*, 2004), text hierarchy is divided using semantic and structural analysis. Each hierarchy represents a sense group and thus the text now becomes a set of sense groups. As shown in Formula (1):

$$T_i = \{C_1, C_2, \cdots, C_m\} = \{[(d_1, w_1),(d_2, w_2)...,(d_k, w_k)], [(d_{k+1}, w_{k+1}),...,(d_l, w_l)],[(d_{l+1}, w_{l+1}),...,(d_n, w_n)]\} \quad (1)$$

where, $k < l < n$ and C_i is the sense group. Each sense group contains the words and the corresponding weights of words. Part-of-speech choice is performed on the words according to their contribution to text categorization. For effective words after part-of-speech choice, the concepts are extracted based on HowNet semantics and the words are mapped into the concept space. Then the sense group and the text can be represented as in Formula (2) and (3):

$$SenseGroup = \{(t_1, w_1),(t_2, w_2), \ldots ,(t_h, w_h)\} \quad (2)$$

$$T_i = \{[(t_1, w_1),(t_2, w_2)...,(t_v, w_v)],[(t_{v+1}, w_{v+1}),..., (t_j, w_j)],[(t_{j+1}, w_{j+1}),...,(t_s, w_s)]\}, v < j < s \quad (3)$$

where, w_i is the weight of concept t_i. The first 20 concepts arranged according to their weights are selected as effective concepts. After dimensionality reduction of sense groups, the texts are stored by text category and then the category sense-group library is obtained. Suppose the category sense-group library is SGC, then SGC = {C_1, C_2, …, C_r}, where C_i is a sense group, with each sense group C_i having *n* characteristic values and, j = 1, 2,… as shown in Formula (4):

$$C_i = \{(t_i^1, w_i^1),(t_i^2, w_i^2)...,(t_i^n, w_i^n)\}, i = 1, 2, \cdots, r \quad (4)$$

With the sense-group category library obtained, the same procedures of sense-group extraction are repeated for text T_i and the sense groups are represented by vectors. An appropriate categorization algorithm is identified between the text to be categorized and category sense-group library. A quantitative relation that can be recognized by the computer is used to determine the category of subject. That is, a mapping relation f is identified so that for $\forall T_i$, we have $T_i \xrightarrow{f} C_i$. By this approach, the text is categorized while saving much time of manual categorization. This approach makes possible information processing and collection.

MAIN MODULES AND THE TEXT CATEGORIZATION ALGORITHM

Syntactic analysis module: It is theoretically believed that the varieties of sentences are infinite, but the types of sentences are finite. Any sentence is classified as basic sentence type or its combination. The major task of syntactic analysis module is to analyze the structure of the sentence and to identify the sentence type. Weights are assigned to sentences according to position of the sentence in the text, the degree of influence of syntax on general idea of the text and the key points of understanding clauses contained in the complex sentence. This process is crucial for the selection of concept features in establishing category sense-group library.

Sentence type classification: Automatic chunk segmentation is used to classify sentence types as well as extract and identify syntactic structure and functional structure on the higher layer. The existing automatic chunk segmentation defines the chunk category from the perspective of syntactic concept (Li *et al*., 2003). By incorporating semantic concept into the definition of syntactic chunk, the grammatical rules are refined and collocation of the structures is constrained. In this way, the grammar and semantic are closely related. In this study, chunk category is divided into two layers, namely, phase element which is grammatical and functional elements which is semantic. Phrase elements include the common phase types: Adjective Phrase (ADJP), Adverb Phrase (ADVP), Location phrase

Table 1: Weight assignment to complex sentences

Sentence type	Processing strategy	Weight assignment
Coordinate complex sentence	Coordinate complex sentence is divided into contrastive sentence and compound sentence. Increased weight is assigned to the second clause in the complex sentence, and reduced weight to the clause in the first part. The weight difference should not exceed 0.5. For compound sentence, the weights of the two clauses are equal.	For contrastive complex sentence, the weight assigned to the first clause is λ_0; $\lambda_0 + 0.5$ is the weight assigned to the second clause.
Progressive, transitional, conditional, cause-and-effect, purposive complex sentence	Increased weight is assigned to the second clause, whereas reduced weight is assigned to the first clause. The weight difference is 0.5.	The weight assigned to the first clause is λ_1; the weight assigned to the second clause is $\lambda_1 + 0.5$.
Explanatory complex sentence	Increased weight is assigned to the explanatory clause; the weight of the remaining part is assigned according to the rules for ordinary sentences.	The weight of the explanatory clause is $\lambda_2 (\lambda_2 > 1)$.
Successive complex sentence	Increased weight is assigned to the last clause; the weight of the remaining part is assigned according to the rules for ordinary sentences.	The weight of the last clause is $\lambda_3 (\lambda_3 > 1)$.
Selective complex sentence	Equal weights are assigned to the first and the second clauses.	No weight adjustment.

(IOC), Noun Phrase (NP), Preposition Phrase (PP), Quantifier Phrase (QP), Verb Phrase (VP); functional elements include Subject (SUB), direct Object (OBJ), Indirect Object (IOBJ) and Preposition Object (POBJ).

The algorithm for sentence type identification adopts sentence type identification strategy based on rule matching. Guided by the syntactic rules of predicate knowledge base and linguistic statistics, sentence type matching is performed for sentence stems. The difficulty in sentence type classification is identification of complex sentences. We use automatic chunk segmentation to divide Chinese complex sentences into 9 categories: coordinate complex sentence, explanatory complex sentence, successive complex sentence, progressive complex sentence, selective complex sentence, transitional complex sentence, conditional complex sentence, cause-and-effect complex sentence and purposive complex sentence (Wen *et al*., 2008). Depending on the contribution of complex sentences to text understanding, we assign variable weights to complex sentences.

Weight assignment: Generally speaking, the title of the text can best reflect the text category. Then the highest position weight is assigned to the title.

After sentence type classification, the text is composed of complex sentences and simple sentences. In light of the degree of the influence of sentence structure and the rules of understanding Chinese complex sentences, we assign different weights to the complex sentences (Table 1). The weights of these words are expressed as the weights of the clause in which the words are located. For repetitively occurring words, their weights are increased by 1. Relevant parameters can be configured in the experiment.

Syntactic analysis module: Semantic analysis module is composed of three steps: semantic hierarchy analysis, part-of-speech choice and concept mapping.

Semantic hierarchy analysis uses statistical semantic analyzer to identify semantic dependency in Chinese sentences (Li *et al*., 2004). The text hierarchies are divided based on semantic dependency, each hierarchy as a sense group. Then the text is composed of the divided sense groups. The elements of sense group are words and the corresponding weights of the words (as shown in Formula: 1).

The theoretical basis for part-of-speech choice comes from the result analysis of a large quantity of text categorization. The general idea of a text is represented by notional words such as verbs, nouns and adjectives; the function words together with high frequency words that appear in various texts are of no use in the categorization. Thus, the function words are filtered out from the sense groups and we will obtain the high frequency stop words. The dimensionality of the vector of characteristic value is reduced, thereby saving the computation time (Xu *et al*., 2008).

Completing the first two steps, the sense groups comprising effective words are obtained. Concept mapping can summarize the semantic information of words as concepts, which effectively removes the adverse impact of synonyms and near synonyms on the categorization. Concept mapping, combining with HowNet semantics, extracts DEF description information of words and represents them as concepts. Then the expression for the sense group has the form of Formula (2). Thus, concept extraction and representation of sense groups are accomplished.

Text categorization algorithm: For text categorization, SVM is used, a population and quick categorization tool at present. SVM is trained with the tagged training text set to obtain the fixed storage of

Table 2: Categorization result comparison

	Classical SVM algorithm			SVM algorithm based on sense group		
Text categorization	Precision	Recall	F1	Precision	Recall	F1
Environmental science	96.762	98.284	97.517	98.243	98.779	98.510
Computer science	96.118	97.201	96.656	97.413	98.056	97.730
Transportation	98.265	98.536	98.400	98.837	98.867	98.852
Education	97.279	98.092	97.683	98.768	96.375	97.556
Economy	94.630	96.257	95.436	95.967	98.134	97.038
Military	97.865	97.013	97.437	97.682	96.815	98.246
Sports	94.713	93.421	94.062	95.126	94.420	94.772
Medicine	95.224	96.732	95.972	97.896	97.734	97.814
Arts	97.300	90.000	93.500	97.300	90.000	93.500
Politics	88.612	97.500	92.811	93.000	98.661	96.400

classified knowledge. The trained SVM is used to categorize the text. The characteristic vectors of the texts to be categorized are input and the fixed storage of classified knowledge is run to obtain the categorization results.

EXPERIMENTAL RESULT ANALYSIS

Experimental assessment approach: In the study of text categorization based on sense group, the categorization is assessed from mainly three aspects: precision (accuracy rate), recall rate and test value of F1.

Precision is the ratio of the number of rightly categorized texts by the categorization approach to the total number of categorized texts, as shown in Formula (5):

$$Precision = \frac{N_{\text{right}}}{N_{\text{total}}} \tag{5}$$

Recall rate is the ratio of the number of texts rightly categorized by the categorization method to the total number of texts that should be categorized, as shown in Formula (6):

$$Recall = \frac{N_{\text{right}}}{N_{have}} \tag{6}$$

The test value of F1 comprehensively considers the two aspects: accuracy rate and recall rate. It is shown in formula (7):

$$F1 = \frac{Precision \times Recall \ 2}{Precision + Recall} \tag{7}$$

Text categorization algorithm: The corpus of Fudan University available on Chinese Natural Language Processing Open Platform is used as the test set. The corpus contains texts of 10 categories: environmental science 264 texts, computer science 280 texts, transportation 250 texts, education 300 texts, economy 325 texts, arts 248 texts, politics 505 texts, military 249 texts, sports 450 texts and medicine 260 texts. We respectively select 200 texts from each category for text training; and 40 texts from each category for text categorization test. We compare the categorization results by classical SVM with those by sense-group-based SVM which is proposed in this study. The parameters in sense-group-based algorithm are configured as follows: sentence weight $\lambda_0 = \lambda_2 = \lambda_3 = 1.3$, $\lambda_1 = 1$ (for parameter meanings, please refer to Table 1; position parameters of sentences: weight of the first sentence is set as 2; weights of the remaining sentences are set as 1. The experimental results are listed in Table 2.

From the above comparison table, we can see that the Chinese text categorization algorithm based on sense group has higher test values. This new algorithm focuses on the features of the Chinese language and combines syntactic and semantic analysis. We obtain increased accuracy rate and recall rate of text categorization. However, due to the lack of distinctiveness in some categories, the accuracy rate of text categorization is affected. Generally, the categories with distinctive contents have higher accuracy rate.

CONCLUSION

Chinese text categorization algorithm based on sense group considers the structural difference between Chinese and English languages. According to the rules of Chinese grammar and semantics, we analyze the sense groups of the trained texts and then extract the sense groups to build category sense-group library. SVM is used for the experiment of text categorization. Chinese text categorization algorithm based on sense group is better adapted to the understanding process of natural language, with more accurately represented texts. Thus, the computer can better understand the contents of the texts. As compared with the conventional categorization approach, the experimental results show that the new algorithm based on sense group has higher precision and its application value is also higher.

REFERENCES

Chen, J. and P. Martha, 2004. Chinese verb sense discrimination using an Em clustering model with rich linguistic features. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, pp: 295-302.

Dai, L., H. Huang and Z. Chen, 2004. A comparative study on feature selection in Chinese text categorization. J. Chinese. Inf. Proces., 18(1): 26-32.

Li, S., Q. Liu and Z. Yang, 2003. Chunk parsing with maximum entropy principle. Chinese J. Comput., 26(12): 1722-1727.

Li, M., J. Li, Z. Wang and D. Lu, 2004. A statistical model for parsing semantic dependency relations in a Chinese sentence. Chinese J. Comput., 27(12): 1679-1687.

Wen, Z., Y. Taketoshi and X. Tang, 2008. Text classification based on multi-word with support vector machine. Knowl. Based Sys., 21 (8): 879-886.

Xu, Y., J.T. Li, B. Wang and C. Sun, 2008. A category resolves power-based feature selection method. J. Soft., 19(1): 82-89

Zhang, H. and Q. Liu, 2002. Model of Chinese words rough segmentation based on n-shortest-paths method. J. Chinese Inf. Proces., 16(5): 1-7.

Zhou, Q., M. Zhao and M. Hu, 2004. A study on feature selection in Chinese text categorization. J. Chinese Inf. Proces., 18(3): 17-23.