• Chinese Keywords Clustering Based on SOM

Yi Wang, Hu Jin
ChengDu University of Information Technology
Wangyi1177@126.com

Abstract

Keyword clustering is useful for text information retrieval, text document classification and so on. This paper introduces an unsupervised method to cluster Chinese keyword by the artificial neural network of SOM (Self-Organized Map). Keywords are encoded into numeric vectors by the similarities of their contextual word sets, which are composed by their neighbor words in the range of phrases. The experimental result shows that words can be clustered on the map according to both of their syntactic and semantic features.

1. Introduction

Keyword clustering is a useful technique for the application of text classification and text information retrieval, which can be implemented by the classical clustering methods such as K-Means or Hierarchical Principle Clustering. Meanwhile, the SOM (Self-Organized Map [1]) is a novel unsupervised clustering method that can form a continuous and visible map of data clusters. The self-organized word map is a kind of self-organized map on which words are distributed according to the similarity of their features. Donald Hindel [2] first proposed a method to classify some English noun-words by their semantic meanings. In his method, a group of predicates were selected as the features of nouns and their co-occurrence with a noun frequencies were used to compose the feature vector of the noun. By using the clustering algorithms such as hierarchical clustering or K-means or others, nouns could be classified into semantic groups that match people's intuition. Later, Timo Honkela proved in his paper [3] that a meaningful self-organized English word map could be formed by just using their neighbor words as their features. On such a map, large linguistic category regions and small semantic regions could be recognized.

The studies of Chinese word clustering are almost based on such two methods. In paper [4], Y. Wen et. al. proposed a method to use the co-occurrence information of adjectives and nouns to classify them simultaneously. M. Qing et. al. presented a method to form the semantic maps of Chinese nouns with the same information used in Y. Wen's study. These studies were mainly focused on the semantic classification of Chinese noun words. The syntactic classification has not been considered.

In this paper, we propose a method to form Chinese word maps on which words, including nouns, verbs, adjectives and adverbs, are distributed according to their syntactic features as well as their semantic meanings. Such maps could be useful for Chinese grammatical parsing based on the point of view that one of the major obstacles of Chinese grammatical parsing is that the immigrated word categories can not accurately reflects the syntactic features of Chinese words. To more accurately embody the syntactic features of words, a Chinese tree bank corpus was used in the experiment as the source of retrieving the syntactic information as well as the semantic information of words. On the generated maps, the distribution of words obviously reflects their similarities both in syntactic and semantic features.

2. The introduction of SOM

The neural network of the self-organized map has two layers: the *input layer* and the *competitive layer*. Every neuron on the competitive layer is connected to all neurons on the input layer with different weights. The learning data are presented to the input layer and stimulate neurons on the competitive layer to be active in different level of activity, which then compete by their activities and the most active neuron would be the winner. The winner, as well as its neighbors in a special radius, modifies their connection weights to be more similar with the pattern of the input data. After a number of iterations of learning, a low-dimensional map that maintains the topologic structure of input data space can be formed. Thus, the competitive layer is

978-0-7695-3304-9/08 $25.00 © 2008 IEEE
DOI 10.1109/ICNC.2008.927

also known as the *mapping layer*. More details about the algorithm in SOM can reference to paper [1].

3. Self-Organized Chinese word map

To form a self-organized word map, the context information of keywords should be retrieved from the corpus and be encoded into numeric vectors to be used as the learning vectors of SOM. In bi-gram language model, the context of a word consists of its predecessor and successor word: {predecessor, successor}. Paper [3] used the quasi-orthogonal random vectors of contextual words to compose the numeric contextual vectors of keywords, and the average vector of all the context vectors of a keyword was used as the learning vector to improve the learning efficiency. Our experiment with the same encoding method to form Chinese word map, however, was unsuccessful. Words distributed on the map irregularly. We regarded that there were two major problems in this method. Firstly, two words neighbored in a sentence do not necessarily imply that they are semantically or syntactically associated. Using improper context words may cause a lot of noise in the learning vectors. Secondly, since Chinese words have no morphologic difference between their different linguistic categories, the contexts of a word in different syntactical situation would be mixed together in calculating the average learning vectors such that they can hardly reflect the different syntactic features of words in different situations. To overcome these problems, we adopted two methods that would be introduced in section 3.1 and 3.2.

3.1 Differentiating contexts

To differentiate the contexts of a word in different situation, a letter that represents the linguistic category was appended at the end of a word. In such a way, a word can be differentiated morphologically according to its different categories in different sentences such that its contexts in different situation can be differentiated either. For instance, the word “发展”(develop) would be changed to “发展 n” and “发展 v”, representing its noun form and verb form respectively.

One thing to be mentioned is that such a modification of words does not necessarily means that their categories are preseted. The purpose of such a modification is just to differentiate the contexts of a word in different situation. Words would be mapped into proper regions according to their contextual features rather than their morphological features. In fact, in our experiment, the word “完全”(thoroughly) that was miss-tagged as adjective in the corpus was mapped in the range of adverbs correctly.

3.2 Retrieving contexts in the range of phrases

The essential of using the neighbor words as the feature of the keywords is that the association of words in corpus implies their special syntactic and semantic relationship, which in turn implies the special semantic and/or syntactic features of keywords. Such an assumption, however, is not always the truth. Sometimes, two words are neighbored in a sentence without any syntactic or semantic relations. For example, in sentence “我们彻底解决了这个问题”(we have thoroughly resolved this problem), the two words “我们”(we) and “彻底”(thoroughly) are neighbored without any syntactic or semantic association.

To eliminate such type of noise data, we use so-called *centralized phrases* instead of sentences as the contexts of keywords. A phrase is a comparatively integrated semantic and syntactic chunk in which words are closely associated. By using phrases as the contexts, the invalid neighborhood of words between two neighbored phrases could be eliminated. On the other hand, since a phrase may consist of one or more sub-phrases, the neighborhood of words between sub-phrases might be invalid either. To overcome this problem, all phrases should be centralized before they are used to be the context of keywords. The *centralization* of a phrase is to retrieve the central word of each of its sub-phrases to compose a new phrase. For example, the phrase “我们热爱伟大的祖国”(“we love great homeland” in word-word translation) contains a sub-phrase “伟大的祖国”(great homeland). The semantic and syntactic successor of the word “热爱”(love) should be the word “祖国”(homeland) rather than “伟大”(great) which is the formal successor in the sentence. To find its real successor, we have to retrieve the central word “祖国”(homeland) from the sub-phrase “伟大的祖国”(great homeland). So-called “central” because the sub-phrase is decoration-central structural and the word “祖国”(homeland) is at the central part of the structure. After centralization, we obtain a new phrase “我们热爱祖国”(“we love homeland”), in which the successor of the word “热爱”(love) is the word “祖国”(homeland). More over, the sub-phrase “伟大的祖国” can be used as the context of the word “祖国” as well, that means we can use the corpus more efficiently as well as more accurately.

Because sub-phrases may contain more deep sub-phrases, the central part of a phrase might be a sub-phrase and its central parts need to be retrieved either. That means, the centralization process is recursive and it would not cease until the central words are retrieved. In other words, the centralization process is a bottom-top operation, in which the central words of the deepest sub-phrases would be returned first to centralize their direct parent phrases, and again, the central words of these centralized parent phrases would be returned to centralize their parent phrases, and so on. Each centralized phrases in this process should be preserved as the contextual source of keywords.

Obviously, only decoration-central structural phrases have the so-called central part. In any language, however, there are a lot of other types of phrases which do not have the central part. The centralization of such types of phrases would be different. For example, in TCT tree bank [5], a sentence structure may be tagged as [DJ NP [VP v n]], in which the subject and the predicate is separated into two levels and their neighborhood cannot be found in phrase range. To resume such neighborhood, the verb phrase (VP) should be replaced by its sub-phrases to transform the DJ phrase into one level structure. Such an operation is called *Expansion*. Another special transforming method is *Replacement*, which should be taken in such situation, e.g., a principle clause contains a subordinate clause. The replacement operation would replace the subordinate clause with a pre-defined word "DJ", which would be the successor word of the predicate in the principle clause such that reflects the syntactic fact that this special predicate can be followed by a subordinate clause.

After preprocessed by the methods mentioned above, the original corpus would be transformed into a new corpus that contains only one-level structural phrases. The context of keywords would be retrieved from within the new corpus.

3.3 Encoding the context

The contexts of keywords are needed to be encoded into numeric vectors to be used in the learning of SOM. In traditional English word map, quasi-orthogonal random vectors were used to encode the context words. Our experiments have proved that such encoding method is unsuitable for Chinese word map. Instead of using quasi-orthogonal vectors, we use the similarity degree of keywords' contexts to form their numeric vectors.

Suppose the predecessor word sets of keyword K_i and K_j are $\{< w_i, n_i >\}$ and $\{< w_j, n_j >\}$ respectively, in which the symbol *n* indicates the number of a predecessor word, and $\{< w_{ij}, n_{ij} >\}$ is the intersection set of these two sets, $n_{ij} = \min(n_i, n_i)$, the similarity degree L_{ij} between them can be calculated by:

$$L_{ij} = \begin{cases} \dfrac{C_{ij}}{C_i + C_j - C_{ij}} & ,i \neq j \\ 1 & ,i = j \end{cases} . \qquad (5)$$

Where, C_i, C_j and C_{ij} are the count of words in $\{< w_i, n_i >\}$, $\{< w_j, n_j >\}$ and $\{< w_{ij}, n_{ij} >\}$, which can be calculated by $C_i = \sum n_i$, $C_j = \sum n_j$ and $C_{ij} = \sum n_{ij}$ respectively. The similarity degree of the successor word sets is the same.

The feature vector of a keyword is composed by its similarity degrees to all other keywords. Suppose the keyword set is $\{K_1, K_1, \ldots, K_n\}$, the feature vector of K_i would be:

$$V_i = \{L_{i1}{}^{prev}, L_{i1}{}^{next}, L_{i2}{}^{prev}, L_{i2}{}^{next}, \ldots, L_{in}{}^{prev}, L_{in}{}^{next}\} \qquad (6)$$

4. Experiments and resultant analysis

4.1 The Corpus and its preprocess

The TCT tree bank [7] was used in our experiment and several preprocesses were taken.

(1) Transforming the Corpus into centralized phrase corpus. The TCT tree bank has 31970 grammatically tagged sentences. After centralization (see section 3.2), we got a new corpus containing 610723 centralized phrases.

(2) Selecting keywords to be mapped. The 100 most frequent words in the corpus, except the punctuations and some very common used words such as "的", "地", "之" etc., were chosen as the keywords that would be presented on the map.

(3) Forming learning vectors of these keywords. Firstly, the contexts of these keywords were retrieved from the centralized corpus and the similar degrees between their neighboring sets were calculated to form the feature vector of keywords, which, according to the method described in section 3.3, had 200 dimensions.

4.2 Learning process in experiments

A square lattice of 16 by 16 was used as the mapping layer of the SOM in the experiment. The learning vectors were presented to the SOM in random order. The learning iteration was set to 20000 times, the initial learning radius and initial learning ratio is set to 8 and 0.5 respectively. After learning, the learning vector of each keyword was presented to the map one more time to find its BMU on the map and tag the BNU with its word string. Since a single unit might be the BMUs of more than one keyword, it could be tagged with multiple word strings.

4.3 Experimental results

The result of the experiment was presented in Figure 1, on which curves were drawn to distinguish different regions that could be recognized. The major regions marked by thick curves reflect the lexical categories of words. The left top corner of the map is the region of verbs, the right-up corner is the region of adverbs, the right middle part of the map is the region of adjectives and the bottom part of the map is the region of nouns.

Unlike others, two verbs "起来 v" and "出来 v", which were circled on the map, are located outside the verb region. After checking their context words, we found that these two verbs have special syntactic feature. They are often used to attach to other verbs and act as the function of some prepositions in English like *up* or *out*. For example, "站起来" can be translated into English as "stand up", where "起来" has the same role of the preposition word "up", or "做出来" can be stated as "work out", here "出来" has the same role of the word "out". So the special location of these two verbs does make sense.

Similarly, three adjectives "完全", "一般", "基本" were mapped outside the region of adjectives. In Chinese, the word "完全" means entire or thorough or absolute, and may be an adjective or an adverb. In the corpus used in our experiment, it is most used to decorate verbs and thus should be tagged as adverb. Unfortunately, it was mistakenly tagged as adjective in the corpus. On the resultant map, according to its contexts, it was correctly mapped into the region of adverbs.

The reason that cause the other two adjective words "一般"(general) and "基本"(basic) being mapped outside the adjective region is more or less the same except that their occurrences as an adjective or as an adverb in the corpus are comparable, which means their contexts in different syntactic situations were mixed and that made them miss-located on the map.

More interesting, in the region of nouns, words were clustered into semantic sub-regions (marked by thin curves in figure 1) that match people's intuitions. Such sub-regions include:

the sub-region of the concept PEOPLE，consisting of words：老师(teacher), 学者(scholar), 教师(teacher), 群众(crowd), 农民(farmer), 朋友(friend), 母亲(mother), 学生(student), 人才(talent), 领导(leader);

the sub-region of the concept LOCATION, consisting of words: 北京(Beijing), 日本(Japan), 美国(USA), 欧洲(Europe)、上海(Shanghai)、英国(England);

the sub-region of the concept SUBJECT, consisting of words: 经济学(economics), 地理学(geography), 图书馆学(library science), 社会学(sociology), 哲学(physiology);

two sub-regions of the concept INSTITUTION, one containing words: 部门(department), 公司(company), 单位(institution), 企业(enterprise) and another containing words: 机构(institution), 组织(organization), 大学(university), 学院(institute), 集团(group), 委员会(committee), 医院(hospital), 机关(department);

two sub-regions of the concept METHOD, one containing words: 出路(way), 做法(approach), 手法(approach) and the second containing words: 方法(method), 措施(measure), 手段(means), 途径(approach), 基地(base), 方针(guideline), 路线(guideline).

Besides these, words in other regions have similar meanings, such as "地区"(region) and "国家"(country); "政策"(policy), "制度"(rules) and "原则"(principle), etc.

To evaluate the effect of our encoding method, we also conducted an experiment using the traditional encoding method proposed in [2]. On the map formed in this experiment, the syntactical regions can hardly be recognized except the adverb region. As the space limitation of the paper, the map in this experiment would not be presented. Table 1 shows the measurements of the two experiments by the meanings of the word's syntactic class.

Table 1 Experimental Results

	Total words	Corrects	Accuracy
SRE[1]	100	43	43%
PRE[2]	100	96	96%

[1] SRE: Sentence Ranged Encoding

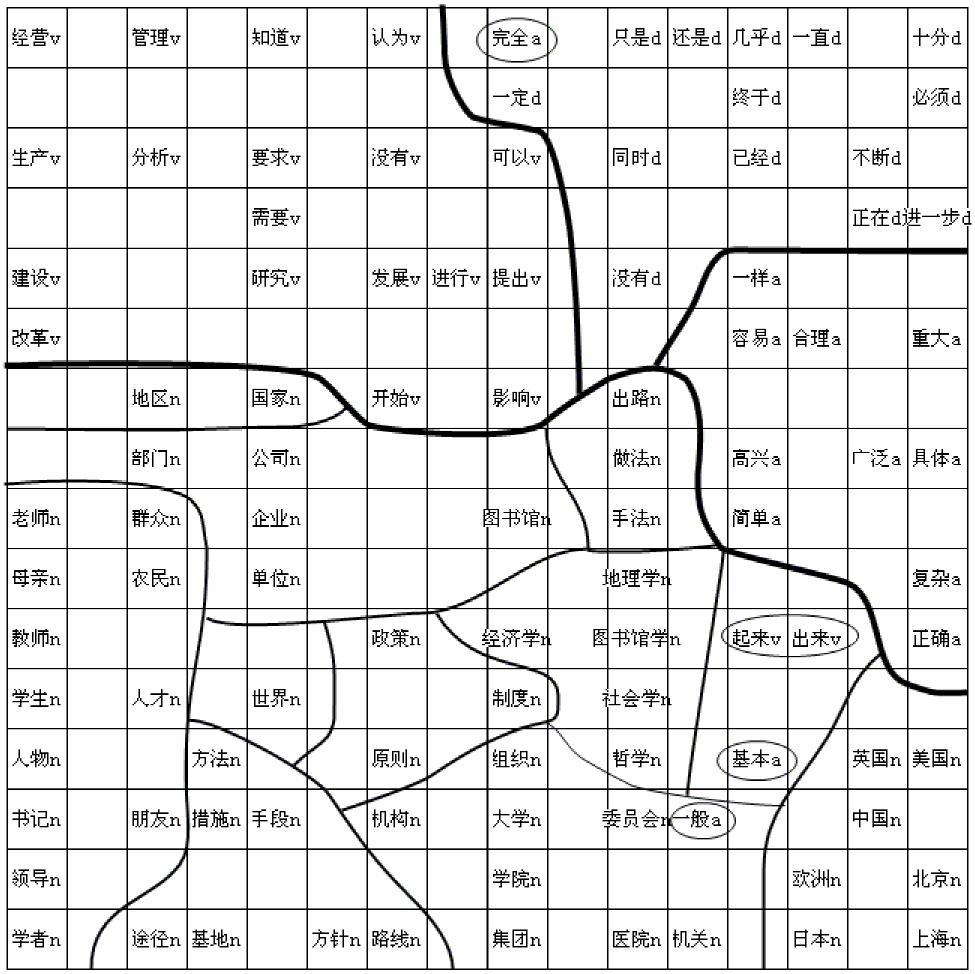
[2] PRE: Phrase Ranged Encoding

5. Conclusion

In this paper, we proposed a method to form self-organized Chinese word map by retrieving the contexts in the range of phrases and encoding them with their mutual context similarity degree. The result of the experiment shows that keywords can be clustered on the map by their syntactic and semantic features implied by their contexts. Such sort of maps, or word classification, could be helpful for the grammatical parsing for Chinese text and for such text information retrievals that consider more semantic information of words.

6. Acknowledge

This work is supported partially by Science and Technology Foundation of Sichuan Educational Department(07ZC008) and partially by Doctoral Start-up Foundation of Chengdu University of Information and Technology(KYTZ200703), and partially by Special Science Foundation of Chengdu University of Information and Technology(CRF200607).



经营v		管理v		知道v		认为v		完全a		只是d	还是d	几乎d	一直d		十分d
								一定d				终于d			必须d
生产v		分析v		要求v		没有v		可以v		同时d		已经d		不断d	
				需要v										正在d	进一步d
建设v				研究v		发展v	进行v	提出v		没有d		一样a			
改革v												容易a	合理a		重大a
		地区n		国家n		开始v		影响v		出路n					
		部门n		公司n						做法n		高兴a		广泛a	具体a
老师n		群众n		企业n				图书馆n		手法n		简单a			
母亲n		农民n		单位n						地理学n					复杂a
教师n						政策n		经济学n		图书馆学n		起来v	出来v		正确a
学生n		人才n		世界n				制度n		社会学n					
人物n			方法n			原则n		组织n		哲学n		基本a		英国n	美国n
书记n		朋友n	措施n	手段n		机构n		大学n		委员会n	一般a			中国n	
领导n								学院n					欧洲n		北京n
学者n		途径n	基地n		方针n	路线n		集团n		医院n	机关n		日本n		上海n

Figure 1 Chinese word map based on phrase contexts

References

[1] T. Kohonen, Self-Organizing Maps, Springer, 2^{d} Edition, 1997.

[2] D. Hindle, Noun classification from predicate-argument structures, proceedings of ACL, 1990, 268-275.

[3] T. Honkela, V. Pulkki ad T. Kohonen, Contextual relations of words in Grimm tales analyzed by self-organizing map, Proceedings of the ICA-95, volume 2, pages 3-7, 1995.

[4] Y. Wen, et al., Clustering Of Chinese Adjectives-Nouns Based on Compositional Pairs, Journal of Chinese Information Processing, 2000, 14(16).

[5] Q. Zhou, W. Zhang, Construction of Chinese Tree Bank, Journal of Chinese Information Processing, 1997