The automatic recognition based on the grade relationship between words of clustering

Juxiang Hu[1], Xueqiang Lv[1], Liping Xu[2]

[1]Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing, 100101, China

[2]Beijing Research Center of Urban System Engineering, Beijing, 100089, China

Keywords: Automatic identification of hierarchy relationship; Thesaurus; Correlation concept space; Co-occurrence analysis; clustering

Abstract. At present, how to use computer technology to automatically identify the relationships between thesaurus of equivalent, grade and related semantic relationships automatically is a key point and also a difficulty for the automatic construction of thesaurus. This paper introduces three typical of concept about hierarchical relationship identification methods: which base on co-occurrence statistics, distribution similarity calculation and syntactic pattern matching, respectively. This paper proposed novel words hierarchical relationship identification method combined co-occurrence statistics and distribution similarity calculation, which demonstrated to be feasible and effective.

Introduction

The rapid development of the network, brings the explosive growth of information resources, provide convenience for people at the same time also make people come to realize that being “overwhelmed” in the ocean of information, which makes us realize to be overwhelmed by the excessive information. Therefore, to obtain the required information accurately and efficiently remains a problem to be solved. "Natural language vocabulary" is a variety of vocabularies that contain natural language elements, or are used to natural language applications and information retrieval according to Zhang[1]. "Natural language thesaurus" refers thesaurus that contains natural language component, which automatically identify the equivalent, grade and correlation relationships between words and construct a vocabulary based on natural corpus by using natural language processing technologies, including pattern matching, co-occurrence analysis and cluster analysis.

There are three kinds of semantic relations between thesauri: equivalent relationships, hierarchical relationships and correlation relationships. To construct thesauri, there are two key steps: vocabulary collecting and identification of semantic relations between vocabularies, It is critical to identify the equivalent, grade and related relationships between vocabularies. In this paper, we realized automatic identification of hierarchical relationships between the vocabularies by combining the co-occurrence analysis technology and words clustering analysis method on the basis of related concept space.

Automatic identification of hierarchical relationship between words

Hierarchical relationships in the thesaurus is in the form of hyponymy, represented by symbols "genus (S)", "sub (F)" and "family (Z)". The thesauri indicate the hierarchical relationships between vocabularies, enlargement or shrinkage has very important significance, can satisfy the need for retrieval of generic retrieval and characteristics. Typical methods for the identification of the concept of hierarchical relationships are as follows [2]:

i)Co-occurrence statistics recognition method based on the concept of hierarchical relationships: this method determines the dependencies between two words according to their co-occurrence frequency. Forsyth and Rada choose high-frequency words as hypernym because of the broad meaning, and choose the low-frequency words as hypernym for the narrow meaning. They designed

a recognition algorithm[3] for hierarchical relationship based on word frequency statistics. Firstly vocabularies are divided into different levels according to frequencies with high frequency words in high-grades and low-frequency words in the lower levels, then the similarities between adjacent level words are calculated from low levels to high levels, and the most similar words are built to the relationship of hyponymy.

ii)Distribution similarity calculation method: the assumption is that the more similar their circumstances are, the greater semantic similar the words are in the context. Many researchers demonstrated the assumption theoretically and experimentally. Distribution similarity calculation method calculates the similarities between two words based on their backgrounds in the context, puts words into groups using clustering method according to the semantic similarities, and identifies hierarchical relationships. In Brown[4] and other studies, each word $T_i$is represented by other words W_i existing in the same article .the context backgrounds are showed by the vector $V(T) = < T_1, W_1 >, < T_2, W_2 >, \dots < T_k, W_k >$ and then the distance between word pairs are calculated using average mutual information formula.

iii)Concept hierarchy recognition method based on syntactic pattern matching[5][6]: the method summarizes typical syntactic patterns and recognizes the relationship between words according to the language style characteristics. The assumption is that the language in the context indicates concept hierarchical relationships of syntactic patterns, such as: ..... Contains (including)..., Belongs to ....., et al.

We constructed hierarchical relationships between words using the "bottom-up" model. It indicate that semantically related words often appear in context simultaneously, we concluded that: In terms of the co-occurrence of a word is often characterized as a vocabulary word construct feature vectors, then the characteristics of the two words vocabulary overlap degree is higher, the higher the semantic similarity. Therefore, co-occurrence analysis techniques can be used to calculate the semantic relevance between words and construct the word "correlation concept space" and then to extract the first K words that are most related to T to construct characterized vector $V(T) = < T_1, W_1 >, < T_2, W_2 >, \dots < T_k, W_k >$, where, T_i is the word associated with the word T, and $W_i$represents the weight of the co-occurrence. Words are classified into clusters depend on subject category using the “hierarchical clustering method”, and then identified hierarchical relationship between the words in the clusters using level recognition algorithm.

Algorithm and experiment

We selected 318 articles of news as the experimental corpus, which were 5 to 10KB published on People's Daily in 2002, and we got 790 words for clustering and hierarchical relationship recognition experiments after filtering processes.

Co-occurrence analysis, concept structure space. In this paper, a single press releases as the co-occurrence window. DICE measure is operated to improve the correlation calculation, and modified adjustment calculation results of DICE measure according to the size of the co-occurrence window. The calculating formula is [7]:

$$W(T_i, T_j) = \frac{2 \times tf(T_i T_j)}{tf(T_i) + tf(T_j)} \times WeightingFactor(T_i, T_j) \quad (1)$$

Where,$W(T_i, T_j)$is the co-occurrence weight of $T_i$andT_j, $tf(T_i T_j)$is the co-occurrence frequency of $T_i$and$T_j$in corpus, and $tf(T_i T_j)$represents the frequency of $T_i$in corpus,

$$WeightingFactor(T_i, T_j) = \frac{\min(length(d_i))}{\sum_{j=1}^{k} length(d_j) / k} \quad (2)$$

is the adjustment factor, tuning the calculation results according to the co-occurrence window: there is a weaker link between words in long literature compared to short documents. $\min(length(d_i))$ is the minimum length of word T_i and T_j in co-occurrence

corpora, $\sum_{j=1}^{k} \text{length}(d_j) / k$ is the average length of co-occurrence corpora, and K is the number of co-occurrence corpora. By calculating the co-occurrence of correlation degree between each word, we can construct a "correlation concept space": point to words, co-occurrence weight as the edge weights of undirected graph.

Word clustering. Selecting the first K words that are most related to word T to construct feature vector: $V(T) = (< T_1, W_1 >, < T_2, W_2 >, \ldots, < T_k, W_k >)$, where T_i is the word related to T, W_i is the co-occurrence weight of T and T_i in concept space. To calculate the semantic similarity between words using cosine similarity algorithm of vector space model, and considering the two characteristic vector simultaneously. If the same key words increase, the feature vector is more similar, so adding a factor to the formula:

$$\text{Sim}(T_1, T_2) = \frac{\sum_{i=1}^{k}(W_{1i} \times W_{2i})}{\sqrt{(\sum_{i=1}^{k} W_{1i}^2) \times (\sum_{i=1}^{k} W_{2i}^2)}} \times \frac{k+n}{k} \quad (3)$$

Where, K is the dimensions of the feature vector; n is the number of feature vector of the same word; W_{1i} represents the value of feature vector of term T_1that is dimension, and $\text{Sim}(T_1, T_2)$ represents the semantic similarity of the word $T_1$and T_2.

This article uses hierarchical clustering algorithm “bottom-up” for the word cluster, firstly to separate each word as a cluster, and then to calculate the distance between each clusters and merger the minimum distance cluster that is the most similar until all words combine into one big cluster. To determine the right feature dimension K and threshold D through the comparison, to compare three methods of hierarchical clustering: simply connected, whole Unicom, average Unicom, and to select the best method.

Hierarchical relationship between word recognition. Hierarchical relationship between word recognition, learn the method of Forsyth and Rada and improve. Analysis of existing Chinese thesaurus, the word in which the level of the hierarchy, with the following two factors: word frequency. The higher the frequency is, the greater the possibility of becoming hypernym. The longer term, its meaning is narrow, become the greater the probability of a word. Forsyth and Rada methods only consider the word frequency factors, this paper to improve them, considering the word frequency and word length, a hierarchy:

$$H(T_i) = \log \frac{tf(T_i)}{len(T_i)} \quad (4)$$

Where, $H(T_i)$ is the Grade coefficient of wordT_i, $tf(T_i)$ express the frequency of wordT_i, $len(T_i)$ express the length of word. According rank coefficient term plan to cluster within each grade in recognition of the word cluster within its upper and lower relationship, the algorithm process is as follows:

Step 1: Determine the number of levels, the word within the clusters classified by rating factor to each word stage; rank high coefficient word in the high word-level, the highest level for the word L_0, the rest followed L_0, L_2, ...

Step 2: Produce the hyponymy relation in between adjacent word level. Extract one word T from the word gradeL_i, calculate every word similarity between word T and word gradeL_i, and take the largest similarity of words as words T hypernym. Go on extracting one word from the word gradeL_i, until all of the words to build up a relationship of Hyponymy.

Step 3: Determine whether to reach the bottom, the end is, or continue to perform the operation of step 2.

Analysis of experimental results

The concept of co-occurrence analysis of spatial structure results.

Table 1: The concept of space

1(word/speech)	2(word/speech)	Correlation
economy/n	construction/vn	0.82572615
Relieve /vn	become profitable/vn	0.42857143
construction /vn	develop /vn	0.8952381
economy /n	develop /vn	0.87323946
…..	……	……

Hierarchical clustering results are as follows:

Table 2: Comparison results of hierarchical clustering algorithm

CA	T	WCN	NCCS	MNCW
SCA	0.1	102	85	657
	0.2	523	463	64
FCA	0.1	367	181	17
	0.2	610	521	11
ACA	0.1	319	155	22
	0.2	571	491	19

Where, CA is Clustering algorithm, SCA is Simply connected algorithm, FCA is Full connected algorithm, ACA is Average connected algorithm, T is threshold, WCN is Word cluster number ,NCCS is Number of clusters containing a single word, MNCW is Maximum number of clusters of words.

For the thesaurus, the distribution of word clusters generally more uniform and less big words clusters, also on behalf of good clustering effect can be divided more evenly distributed clusters. The results from cluster analysis can be seen: The average connectivity hierarchical clustering algorithm, the threshold value of 0.1 is better.

Table 3: The result of clustering experiment

No.	Num	The words in the word cluster
1	2	Administrative area/n 89||Hong Kong Special Administrative Region/n 44
2	3	apple/n 28||fruit tree/n 27||fruit/n 17
3	3	resource/n 263||industry/n 142||productivity/n 137
4	6	literature and art/n 111||art/n 86||composition/n 47||longhair/n 37||crosstalk/n 21||TV station/n 18
5	15	News/n 112||incident/n 92||network/n 81||accident/n 66||information/n 56||channel/n 53||Public opinion/n 50||media/n 38||forum/n 35||hot pints/n 35||newspaper/n 23||comment/n 19||net friend/n 19||Website /n 16||internet /n 15
.	…	…….

The experiment result of hierarchy relationship recognition.

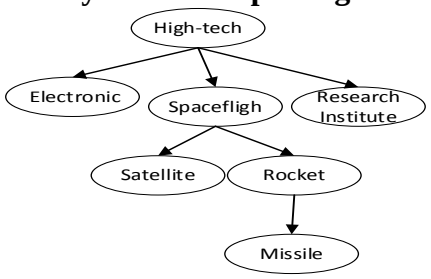

Fig.1 The relationship of hierarchy

Through experiments, the following conclusions can be draw: through the calculation of correlation between words co-occurrence analysis, can be identified without literal similar characteristics of word similarity; On the basis of the structure "related concept space", using hierarchical identification method, basically can distinguish the word that express different themes category, generated word cluster distribution is uniform, within the cluster similarity between each word is higher; Using grade recognition algorithm, basically word within a cluster can be classified into different grades, and then through artificial determination and adjustment to determine the level of relations between words. According to the obtained clustering word from experiments which is not accurate enough, the identified hierarchical relationship need to manually adjust, can use a large corpus to do experiment, further improve the grade of the recognition algorithm, to get more ideal experimental effect.

Conclusions

This paper propose a combination of co-occurrence statistics and distribution similarity calculation between word hierarchical relationship identification method, overcome the simple according to the grades of the clustering method to extract the disadvantages of vocabulary, according to the setting threshold value was verified the effectiveness of the experiment. Based on word clustering method still exist deficiencies, for example: the performance of the clustering algorithm remains to be further improved, so the next step of work need to be further explored, and the result was further improved.

Acknowledgements

This work is supported by National Natural Science Foundation of China under Grants No. 61271304 and Beijing Natural Science Foundation of Class B Key Project under Grants No. KZ201311232037

References

[1] Zhang Qi-Yu. Positive for the combination of natural language and information retrieval language create conditions, which proposed a large number of natural language vocabulary [J]. Library Journal, 1999(9)

[2] Du Hui-Ping. The automatic identification of concept hierarchy [J]. China Index, 2010 (3)

[3] R. Forsyth, R. Rada. Adding an edge, Machine Learning: Applications in Expert Systems and Information Retrieval[M]. Chichester: Ellis Horwood Ltd,1986

[4] Brown P, Della P S, Della P P, Mercer R. Word sense disambiguation using statistical methods[J]. In Proceedings of the 29th Meeting of the Association for Computational Linguistics(ACL-91),1992

[5] Hearst M A. automated discovery of Wordnet relations[M]. Cambridge: MIT Press,1998

[6] Yousef ABUZIR. Deriving concepts hierarchy[EB/OL].

[7] Du Hui-Ping, Zhong Yun-Yun. Natural language thesaurus automatically building research [M]. Southeast University Press, 2009