Research of Conceptual Relation Extraction Based on Improved Hierarchical Clustering Method

Caiyun Xie* and Junyun Wu

Department of Information Science, Nanchang Teachers College, Nanchang, 330031, China; 2School of Information and Engineering, Nanchang University, Nanchang, 330031, China

Abstract: The main task of Ontology learning is concept extraction and conceptual relation extraction. This paper mainly studies the latter. Conceptual relation consists of taxonomic relation and non-taxonomic relation. It introduces hierarchy clustering method, and uses concept hierarchy clustering method which chooses different clustering standards in each hierarchy to obtain the taxonomic relation. It improves the accuracy of the relationship extraction. For extracting the non-taxonomic relation, this paper uses a extended association rule, this method can get concrete names of relation, and confirms the domain and range. In the end, the paper uses the introduced method of Ontology Learning to constructing a domain ontology in the law. And it completes the implementation of an Ontology-based semantic retrieval system. The final effect of this system application demonstrates that this Ontology learning method is efficient.

Keywords: Conceptual Relation Extraction, Extended Association Rule, Hierarchy Clustering, Ontology Learning.

1. INTRODUCTION

In recent years, more and more researches on Ontology have appeared in computer science. The most widely cited definition of Ontology is proposed by Studer [1], who said " Ontology is a clear formal specification of shared conceptual model." It is used to record the concepts of a particular area and the direct relationship between these concepts, so that these concepts have the clear, formal definition in the case of being shared, and which facilitates communication between the operator and the machines.

So far, the types of ontology construction tools have been more and more, such as: Protégé [2], WebODE [3] and KAON [4] and so on. Appearance of these tools makes the ontology construction convenient for users, but these tools still have flaws, users can only use them to edit ontology, other content such as concept, conceptual relations, constraints, and so need to be entered manually, it is not realistic to build by hand for the larger scale of Ontology. Therefore, it is difficult to construct Ontology manually because of time-consuming and being difficult to update. In order to promote the development and application of Ontology, either automatically or semi-automatically ontology construction and ontology learning techniques have been proposed and become a hot research[5], the main task of ontology learning is concept extraction and conceptual relation extraction.

Relations between concepts need to be extracted after extracting the domain concepts. So it is an important task. Most of existing methods of relation extraction used in the study are for English terms, put them directly for Chinese Ontology is not ideal [6]. The main reason is that the basic unit of Chinese is character, and there is no space between every two words or terms as English words, so they are lack of morphological markers. Hence, it is difficult to find the dividing point between the words. So the term set get by using the segmentation tool may not be accurate, we must take into account that some words which are apart may include compound words.

2. THE OVERVIEW OF CONCEPT RELATION EXTRACTION METHOD

There are mainly two categories of relationship between concepts in Ontology: taxonomic relation and non-taxonomic relation. Taxonomic relation includes upper and lower classification relation (is-a relationship), non-taxonomic relation includes whole-part relationship, related relationship, synonymous relationship, causality relationship, etc. Therefore, these two relations are mainly considered relationship in the process of getting conceptual relation.

There are many kinds of methods to obtain conceptual relations, including the following categories:

1. Method based on the dictionary

The dictionary we are talking about contains language dictionaries and specialized dictionaries, which are a collection of some words. And these words are distributed in a certain order. With the continuous development of language, the dictionaries which can reflect semantics have appeared now. The dictionary what we are most familiar with is Wordnet, generated by a research of Princeton University. Chinese dictionary "HowNet" and "Modern Chinese semantic Dictionary" have more authority. Dictionary-based method extracts the relationship between ontology concepts by looking for synonyms and antonyms words.

2. Method based on vocabulary-Syntactic pattern matching

Pattern in this method mainly refers to syntactic patterns, including some rules. This method mainly uses the way of analyzing, summarizing corpus in related fields. Using this method to get the conceptual relations basically compares the text and the pattern to find concepts organizing pairs of relations which conform to rules. Hearst [7] etc. proposed this method and used in extracting conceptual relations in Ontology in 1992.

3. Method based on concept clustering

Clustering method is based on information-similar criteria; it classifies information under the condition unknowing the information needed to be classified. When extracting conceptual relations, the criteria used is to calculate the distance between the concepts of the semantic level, then the concepts have the minimum distance are gathered into the same cluster by the method of concept clustering. Because the clustering criteria need to be taken into account, the clustering criteria can affect the results relatively.

4. Method based on association rules

Method based on association rules is mainly used to obtain the non-taxonomic relations among concepts. It includes two processes: in the first stage, to find out all pairs of relations with higher frequency from collection of texts; in the second stage, to organize these relations to generate association rules. There is already a lot of research on obtaining conceptual relations using association rules, but most can only confirm the existence of the relations between the two concepts, they can not find out what kind of specific relationship.

5. Mixed method

In practical ontology learning, the above two or more methods are often used to obtain the relations between concepts, we expect to get better experimental results.

3. TAXONOMIC RELATIONS EXTRACTION

Taxonomic relation [8] is similar to inheritance in Object-Oriented Programming, that is, the relationship between the upper and lower bits. Such as "Criminal Law" in the legal field, it belongs to "the law", so they have a direct relationship between the upper and lower, which is Taxonomic relation. "Law" is the upper class, "Criminal Law" is the corresponding lower classes. This paper mainly introduces the concept clustering method for relation extraction, and makes related improvements for incomplete place.

3.1. Overview of Concept Clustering Methods

This paper has briefly introduced the idea of the concept clustering. Concept clustering method is mainly used to find taxonomic relations between concepts, which builds the vector space model for context information of concepts. It calculats semantic similarity through vector space between concepts, and then classification of concepts can be obtained according to the closeness of semantic similarity. The commonly used methods of concept clustering includes hierarchical clustering, plane partitioning method and formal concept analysis (FCA).

1. Hierarchical clustering method

The hierarchical clustering method is to stratify the given data objects, in accordance with the criteria of from the hierarchical data objects stepwise clustering terminates when it reaches a suitable condition. Data objects in the distance calculation method of this paper is by calculating the semantic similarity between the words, and then the concept of clustering. Hierarchical clustering methods include two types: split hierarchical clustering method and condensational hierarchical clustering method. Split hierarchical clustering method is to consider all the objects as a category, and to divide them into several small groups gradually according to certain criteria, until each object become a class or reaching condition of termination. The terminal condition contains that the number of classes reaches a desired result or the distance between the classes reaches a certain threshold. On the contrary, the condensational hierarchical clustering method considers all the data objects as a separate class, and then gets two or more classes combined into one category according to certain criteria, until all objects form a class or condition of termination.

2. Plane partitioning method

The idea of plane partitioning method is to create a division including k numbers, and then to relocate division using an iterative approach between the concepts to generat new division. The difference between this method and hierarchical clustering method is that the former divide the set of documents in horizontal position rather than vertical division. The specific approach is to generate a cluster center containing k numbers, and then to calculate the similarity of each concept and various centers in the seed and find the seed with maximum value of similarity to cluster. Plane partitioning method is often divided into two kinds: K-Means algorithm and K-center point algorithm.

3. Method of formal concept analysis (FCA)

Formal concept analysis method represents the form of background in the field by using binary relations. This method extracts conceptual level from the background which is also called concept lattice. Formal concept analysis is to get hierarchical data through that the concept lattice is structured.

3.2. Conceptual Relation Extraction Based on Improved Hierarchical Clustering Method

This section will further introduce Conceptual relation extraction based on improved hierarchical clustering method. The clustering method is according to the similarity between the concepts, the method of calculating the similarity is as follows.

In this paper, the concepts are formed a "term-document" matrix expressed as $M[m][n]$, where m is the number of concepts, n indicates the number of documents in the professional corpus. For example, $M[i][j]$ indicates the number of term i appearing in the document j. Row of the matrix represents a concept vector, it is expressed as follows:

$$T_i = (M[i][0], M[i][1], \cdots, M[i][n]) \tag{1}$$

Based on the above information, it calculates the similarity between concepts using the cosine similarity. In order to avoid losing the similarity between some of the concepts , this paper introduces similarity in HowNet to adjust the relevant formula, as is shown follows:

$$sim(T_i, T_j) = \frac{simA(T_i, T_j) + \alpha \cdot simB(T_i, T_j)}{2} \tag{2}$$

Where $simB(T_i, T_j)$ is the Semantic similarity between concepts in HowNet; α is an adjustable parameter, the value is usually set as 0.5.

After the preparation work, hierarchical clustering steps are given below:

1. Choose any k concepts as the center of clustering collections, expressed as $(\mathrm{C}_1, \mathrm{C}_2, \cdots, \mathrm{C}_i, \cdots, \mathrm{C}_k)$, which is the initial state of clustering;

2. Use the formula (4.2) to calculate the similarity of each concept with all the cluster center, find the maximum similarity of each concept with each class (ie, where the collection of cluster centers) to merge into it;

3. Do step 2 circularly until the end of the two clustering meet the termination condition.

3.3. Improved Method of Relation Extraction Based on Hierarchical Clustering

As the problem of hierarchical clustering method is to choose the reference value when clustering, that is clustering criteria. If the value of clustering criteria has too low discrimination, the results of the experiments are less accurate. Further, the limitation of this method is that the concept having multiple parent concepts will not be recognized [9]. Therefore, this section can also be used to improve the method, that is to cluster multiplely to identify all the relationships.

Therefore, for the limitation which the selected criteria affect hierarchical clustering result , it gives the corresponding improved method. Specific approach is to make multiple hierarchical clustering, and each process uses different standards. This will improve the accuracy of relation extraction.

Specific steps are as follows:

1. Choose any k concepts as the central concept of clustering collections, expressed as $(\mathrm{C}_1, \mathrm{C}_2, \cdots, \mathrm{C}_i, \cdots, \mathrm{C}_k)$, which is the initial state of clustering;

2. Use the formula (2) to calculate the similarity of each concept with all the cluster center, find the maximum similarity of each concept with each class (ie, where the collection of cluster centers) to merge into it;

3. According to the algorithm given below to calculate the next round of k classes in the cluster center:

(1) Calculate the average similarity of each concept in class i $(\mathrm{i} = 1, 2, \cdots, \mathrm{m})$, assume that there are n concepts totally, using the following formula:

$$asim[i] = \frac{\sum_{k=1}^{n} sim(T_k, C_i)}{n} \tag{3}$$

(2) Identify r concepts which are closest to the clustering center from the class in step (1), r is calculated as follows :

$$r = m * \frac{asim[i]}{\max\_\, asim} \tag{4}$$

Where $\max\_\, asim$ is the maximum value obtained by calculating equation (3).

(3) Calculate the of r concept-vecters , take the concept most similar to the average as the center of the clustering collections of the next round ;

4. Compare the result from the step (3) with the last round of cluster center, if the the degree of difference between the two is greater than a given threshold, continue to perform step 2, otherwise go to step 5;

5. Obtain m classes, end all steps.

This paper uses the improved hierarchical clustering method to extract conceptual relations, specific algorithm flow chart is given in Fig. (**1**):

4. NON-TAXONOMIC RELATION EXTRACTION

Non-taxonomic relation is all the relations except taxonomic relation (upper and lower classification relation), including whole-part relationship, related relationship, synonymous relationship, causality relationship, etc. Compared with taxonomic relation, non-taxonomic relation is more complex relationship, this relations generally extracted by association rules.

4.1. Method Based on Association Rules

Method based on association rules is mainly used in non-taxonomic relation extraction. This method indicates that if two concepts often occur concomitantly with each other within the same field of the document, there is some relationship between them.

The basic idea of association rule mining is to use statistical methods to find the association between two things. Therefore, for the extraction of non-taxonomic relations between concepts, we can make use of association rules. The introduction of association rules is given as follows.

Assume that collection $S = \{I_1, I_2, \cdots, I_m\}$ represents a group of items, collection $T = \{t_i \mid i = 1, 2, \cdots, n, t_i \subseteq S\}$

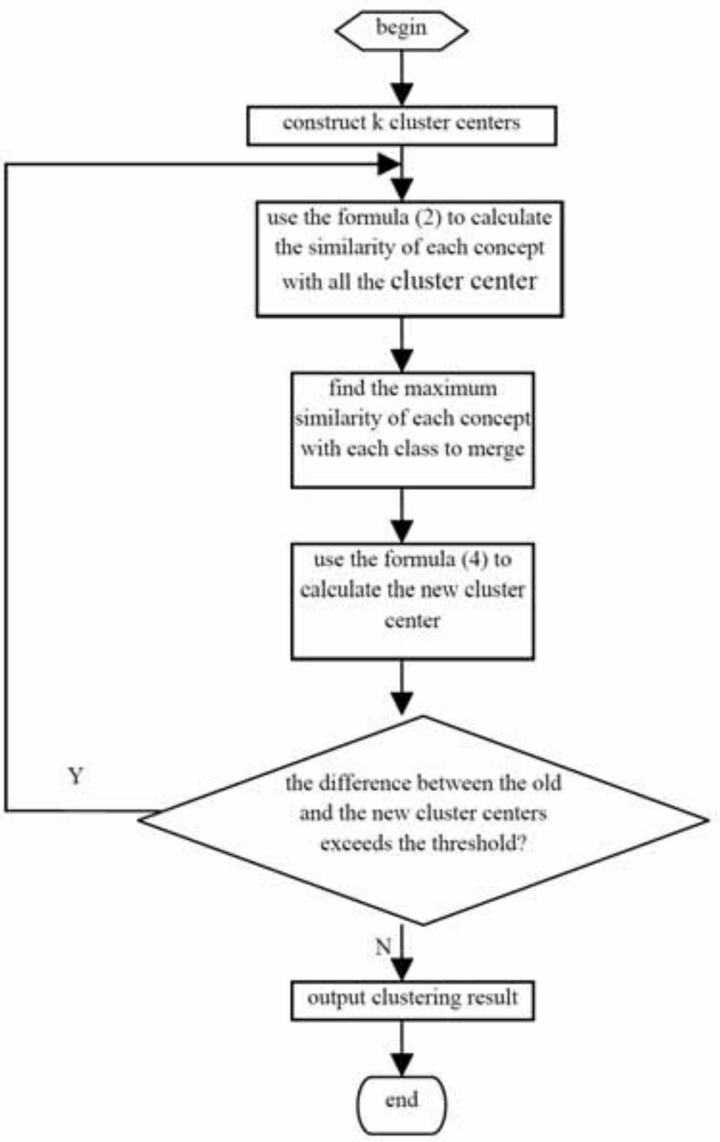
begin
construct k cluster centers
use the formula (2) to calculate the similarity of each concept with all the cluster center
find the maximum similarity of each concept with each class to merge
use the formula (4) to calculate the new cluster center
the difference between the old and the new cluster centers exceeds the threshold?
Y
N
output clustering result
end

Fig. (1). Flow chart of improved hierarchical clustering method.

represents a set of transactions. If the items X,Y are two groups, and $X,Y \subseteq S$ but $X \cap Y = \varnothing$, the definitions and calculation of credibility and support of association rules $X \Rightarrow Y$ are as follows:

Definition 1 Credibility of $X \Rightarrow Y$ is the probability of Y appearing in the premise of X appearing in affairs T, expressed as $confidence(X \Rightarrow Y)$.

$$confidence(X \Rightarrow Y) = P(Y \mid X) = \frac{\left|\left\{t_i \mid t_i \supseteq X \cup Y\right\}\right|}{\left|\left\{t_i \mid t_i \supseteq X\right\}\right|} \tag{5}$$

Definitions 2 Support of $X \Rightarrow Y$ is the probability of X and Y occuring in transaction set T simultaneouly, expressed as $\sup port(X \Rightarrow Y)$.

$$\sup port(X \Rightarrow Y) = P(X \cup Y) = \frac{\left|\left\{t_i \mid t_i \supseteq X \cup Y\right\}\right|}{n} \tag{6}$$

Association rules method means: if and only if rule $X \Rightarrow Y$ reaches the minimum threshold of credibility and support, the rule $X \Rightarrow Y$ is established in the transaction set T.

When taking the association rules applied to non-taxonomic relation extraction, the item sets represents the concepts set of existing non-taxonomic relationship, expressed as $S = \{C_1, C_2, \cdots, C_m\}$. And t_i is a sentence which at least contains one concept from the transaction set. $X = \{C_i\}$ and $Y = \{C_j\}$ are the concepts consisting non-taxonomic relation for association rules extraction. Therefore, according to the association rules, when $confidence(X \Rightarrow Y)$ and $\sup port(X \Rightarrow Y)$ are greater than a given threshold at the same time, there is a strong relevance between these two concepts. This pair of concepts is the non-taxonomic relation that we want to extract.

Table 1. The comparison of conceptual relationship results.

		Accuracy	Recall	F-value
taxonomic relation	before improving	52.4%	45.2%	48.5%
	after improving	54.7%	48.0%	51.3%
non-taxonomic relation	before improving	39.1%	47.2%	42.8%
	after improving	40.8%	49.7%	44.8%

However, extracting the relationship using this method only can determine the existence of a relationship between two concepts, but can not determine what these two concepts are, i.e. particular relation name can not be drawn. So method based on association rules needs to be improved. This paper introduces the extended association rules to get the non-taxonomic relations with concrete names, and give the domain and range where all relations belong.

4.2. Relation Extraction Based on Extended Association Rules

Dependency syntax indicates that verb the core element dominating the other ingredients in sentence, the verb itself is not bound by the other ingredients and other ingredients in two sides of the verb does not have inter-related [10]. To the Chinese text, the main component of the sentence is SVO, the type of predicate is usually verb. According to the dependency syntax, there is no relevance between subject and object with each other. So for this kind of sentence, the verb can be used directly as the name of non-taxonomic relation, the main component of subject is taken as the domain of the relation, the main component of object is the range of the relation.

What describes above is the thought of extented association rules. Before obtaining the specific relationship, we can use association rules to extract the pairs of concepts existing non-taxonomic relation, where the two concepts are likely the subject and object of some sentence. So just find the verb dominating the two concepts, we can determine the relation name. We can give a threshold, if the times of a verb appearing in a sentence which contains a concept pair exceeds this threshold, it is determined that the verb is the relarion name of the concept pair. And the concept appearing in front of the verb is the domain of this relationship, the concept behind it is the range of this relationship.

The algorithm is described as follows:

(1) Extract any two concepts c_1, c_2 that are not analyzed by association rules from extracted set of concepts. If the concepts that are not analyzed can not be found, perform step 6;

(2) Calculate credibility and support of $c_1 \Rightarrow c_2$ according to equation (5) and (6) respectively;

(3) If the credibility and support of $c_1 \Rightarrow c_2$ reache a given threshold, perform step 4;

(4) Calculate the number and co-occurrence frequency of the verbs and c_1, c_2 appearing in the same sentence using statistical method;

(5) If the frequency of a verb reach a given threshold, this verb is the relation name of the concept c_1, c_2, the concept appearing in front of the verb is the domain of this relationship, the concept behind it is the range of this relationship. Complete analysis of all verbs, perform step 1;

(6) End all steps.

5. EXPERIMENTAL RESULTS AND ANALYSIS

This section analyzes the validity of relation extraction algorithms introduced in this paper. The accuracy of experiment (precision), the recall rate (recall) and F-value (F-measure) are three kinds of evaluation index [11-13], they are defined as follows:

Accuracy rate = the number of correct relations / the number of extracted relations * 100%

Recall = the number of correct relations / actual number of extracted relations * 100%

F-value = 2 * accuracy * recall / (accuracy + recall)

This paper extracts taxonomic relation by using the multiple hierarchical clustering based on the extracted concepts. Whether the difference between the cluster center reaches the given threshold, the value is set as 0.001 in this experiment; extracting non-taxonomic relations by using extended association rules also needs to set the threshold for the credibility and support, according to the conclusion of reference [14]: when the minimum support is set as 0.0001 and the minimum credibility is set as 0.05, the accuracy of extracted concepts is the highest rate. So the paper also put this threshold for non-taxonomic relation extraction. Table **1** shows the comparison of the two kinds of relations before and after improving.

Table **1** shows that the accuracy of taxonomic relation is higher, because this paper chooses the clustering criteria many times when using hierarchical clustering. So accuracy is improved. For non-taxonomic relation, the accuracy is lower, because of using the method containing rules, in which the system of rules is not complete. Overall, the accuracy and recall rates are improved.

CONCLUSION

This paper introduces several methods of relation extraction, and then introduces the extraction methods of taxonomic relation and non-taxonomic relation. It introduces hierarchy clustering method, and uses concept hierarchy clustering method which chooses different clustering standards in each hierarchy to obtain the taxonomic relation. It improves the accuracy of the relationship extraction. For extracting the non-taxonomic relation, this paper uses a extended association rule, this method can get concrete names of relation, and confirms the domain and range.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

The project is supported by Foundation of Jiangxi Provincial Education Department (No.GJJ14791) and Jiangxi Provincial Department of Science Technology (No. 20122BBE500049).

REFERENCES

[1] M. Uscholdal, and M. Gruninger, “Ontologies: principles, methods and applications,” *The Knowledge Engineering Review,* vol. 11, pp. 93-136, 1996.

[2] N. F. Noy, R. W. Fergerson, and M. A. Musen, “The knowledge model of protégé2 2000: combining interoperability and flexibility,” In: *Proceedings of the International Conference on Knowledge Engineering and knowledge Management Knowledge Patterns,* pp. 17-32, 2000.

[3] J. C. Arpirez, O. Corcho, and M. Fernandez-Lopez, “WebODE: a scalable ontological engineering workbench,” In: *Proceedings of the Knowledge Capture,* pp. 6-13, 2001.

[4] E. Bozsak, M. Ehrig, and S. Handschuh, “KAON2-towards a large scale semantic web,” In: *Proceedings of the 3rd International Conference on E2-Commerce and Web echnologies,* 2002, pp. 304-313.

[5] Q. Yan, and X. Hongwei, “The research of concept extraction based on web data mining for ontology learning,” *Conputer Development & Applications,* vol. 20, no. 11, pp. 37-39, 2007.

[6] D. Liuling, H. Heyan, and C. Zhaoxiong, “A comparative study on feature selection in chinese text categorization,” *Journal of Chinese Information Processing,* vol. 1, no. 18, pp. 123-126, 2004.

[7] M.A. Hearst, “Automatic acquisition of hyponyms from large text corpora,” In: *Proceedings of the 14th International Conference on Computational Linguistics,* France, 1992, pp. 539-545.

[8] G. Rigau, H. Rodrigues, and E. Agirre, “Building accurate semantic taxonomies from monolingual MRDs,” In: *Proceedings of the COLING-ACL,* San Francisco: Morgan Kaufmann Publishers, 1998, pp. 1103-1109.

[9] Y. Yufei, D. Qi, J. Zhen, and Y. Hongfeng, “Weakly supervised method for attribute relation extraction,” *Journal of Computer Applications,* vol. 34, no. 1, pp. 64-68, 2014.

[10] W. Chun, S. Zhaoxiang, and X. Yuan, “Chinese non-taxonomic relation extraction based on extended association rule,” *Computer Engineering,* vol. 35, no. 24, pp. 63-65, 2009.

[11] S. Zhuting, “Extraction of concept relationship in the process of construction concept maps”, *Journal of Qiongzhou University*, vol. 21, no. 2, pp. 22-27, 2014.

[12] H. Xun, Y. Hongliang, and Y. Yang, “A review of relation extraction”, *New Technology of Library and Information Service*, vol. 239, no. 11, pp. 30-39, 2013.

[13] J. Villaverde, A. Persson, D. Godoy, “Supporting the discovery and labeling of non-taxonomic relationships in ontology learning,” *Expert Systems with Applications,* vol. 36, no. 7, pp. 10288-10294, 2009.

[14] W. Chun, S. Zhaoxiang, X. Yuan. Contrast research of Chinese domain ontology concept hierarchy induction methods,” *Application Research of Computers*, vol. 26, no. 8. 2847-2850, 2009.

Received: October 16, 2014 Revised: December 23, 2014 Accepted: December 31, 2014