Research on Construction Method of Chinese Domain Ontology Based on Relation Extraction

Shanshan Qi¹, Limin Zheng¹,²* and Lu Yang¹,²

¹College of Information and Electrical Engineering, China Agricultural University, Beijing, 100083, China
²Beijing Laboratory of Food Quality and Safety, China Agricultural University, Beijing, 100083, China
*Corresponding author’s e-mail: zhenglimin@cau.edu.cn

Abstract. Domain ontology plays an increasingly important role in the information extraction process and has become a useful tool for knowledge acquisition, knowledge sharing, and modeling. However, the existing domain ontology construction methods are mainly based on manual mode, lacking systematic and engineering methods. This paper proposes a construction method of Chinese domain ontology based on relation extraction, which fully considers the automatic extraction of semantic hierarchical relations, and guarantees the semantic richness and accuracy of relational triples. The construction method of Chinese domain ontology based on relation extraction is applied to the field of food safety events. The constructed ontology has the advantages of reusability and scalability, which proves that the method is logical and operability.

1. Introduction

The amount of information on the Internet has shown an exponential growth trend, advanced search engines often search out a large amount of redundant and complex information, so people cannot quickly, clearly and accurately select the desired content. However, the proposing of the ontology brings an opportunity for knowledge representation and knowledge organization, and can effectively solve the problem of knowledge acquisition to a certain extent. The ontology was originally derived from philosophical concepts and was later introduced into many fields of computer science, such as artificial intelligence, knowledge engineering, and library and information. Ontology in computer science refers to the formal definition of types, attributes, and relations between entities in a domain[1].

The functions of establishing ontology lie in: to reach a consensus on information structure between people and software agents, and to facilitate man-machine dialogue; to share and reuse knowledge in this field; to achieve interoperability between different systems and models in the field[2]. Domain ontology is a professional ontology, which displays the specific concepts and terms in the field through relevant knowledge of the domain, and comprehensively describes the attribute information and category information of these terms, as well as the semantic relations between related terms from different perspectives, so as to provide a comprehensive and clear understanding of domain knowledge. However, at present, there is not a set of mature and complete guidance method for ontology construction research. Building a relatively complete domain ontology requires not only obtaining a large number of concepts and terms but also dealing with the semantic conflicts and ambiguity.
between them. As a result, the construction of domain ontology becomes a bottleneck of knowledge acquisition due to its heavy workload, time-consuming and laborious.

In order to solve the difficulties encountered in the process of domain ontology construction, more and more researchers have carried out deeper exploration. For example, in foreign countries, Cho-Wei Shih et al.[3] proposed a method for exploring the relations between concepts of crystallization, which combines the characteristics of two methods from top to bottom and from middle to outside in the process similar to snowflake crystallization. It implements an ontology construction mechanism, which can automatically mine domain concepts from domain documents, determine the relations between concepts, and construct domain ontology accordingly. However, this method faces the problem of difficulty in initial seed selection, and the constructed result is greatly influenced by seed quality. Yu et al.[4] proposed a semi-automatic domain ontology construction framework based on web crawler, which obtained domain data on the network, and extracted semantic knowledge through linguistic and statistical methods. Finally, the ontology was constructed using the extension-based domain ontology modeling method. The framework can save human resources, but it does not consider the semantic distance between terms, and the accuracy of the results cannot be further ensured. Muhammad Amith et al.[5] first organized a classification scheme of four ontology assessment methods in the existing literature to create a comprehensive classification. They also investigated the biomedical ontology and the latest quality assurance methods used and mapped them to ontology evaluation standards. It is expected that ontology assessment and quality assurance methods will be more widely used in the development life cycle of biomedical ontology.

A large number of domain ontology construction studies have also appeared in China. For example, Tang et al.[6] studied the domain ontology construction based on the thesaurus, and combined the Enterprise, Methontology method and software development model that is “waterfall model”, proposed a method of domain ontology construction based on the thesaurus. But there is no thesaurus can be used directly in most fields, and the construction of the thesaurus requires a lot of manpower, time, and professional knowledge. Wang et al.[7] proposed a Jena-based domain ontology reasoning model for film material, which can well represent the knowledge of film material, not only supports intelligent query of semantic reasoning, but also has high efficiency. However, due to the fact that too many concepts are covered in ontology, the construction process is still not perfect. By investigating the construction of domain ontology in large-scale applications, Fu[8] proposed the method of constructing domain ontology in the context of big data. The constructed health domain ontology has the advantages of reusability and scalability. This research can provide a reference for the rapid and effective construction of domain ontology in the context of big data.

With the development of economy and society, a large amount of knowledge on the network has played the role of information transmission. However, there are some shortcomings, such as information redundancy and disorder, need to be selected manually, not intuitive and not systematic, which make it more and more difficult to meet people’s needs. In the Semantic Web, the semantics of information is described by ontology, and ontology can describe the concepts and their connections in a specific domain. In this regard, we can use the rapidly developing Internet technology to obtain relevant data from the network and then store and process it to build domain ontology. We can present information intuitively, systematically and reliably to users, in order to meet the needs of people for domain information, and solve this problem to a certain extent. Therefore, we propose a domain ontology construction method based on relation extraction and apply this method to construct the domain ontology of food safety events.

2. Domain ontology construction method
Ontology construction is also known as Ontology Learning[9]. At present, there are seven kinds of modeling methods commonly used in domain ontology, and the order of maturity of development is: seven-step method[10], METHONTOLOGY method[11], IDEF5 method[12], TOVE method[13], skeleton method[14], the SENSUS method and KACTUS method[15]. Most of the ontology construction methods have their unique principles and design standards. The core task of domain ontology construction
based on relation extraction is to mine the semantic relations between concepts and concepts, terms and terms from Chinese text. Finally, it can be visually represented by visualization technology or other organizational methods to speed up the information query rate.

Referring to the above methods of domain ontology construction and the specific situation of information extraction, the research framework of the domain ontology construction method based on relation extraction is shown in figure 1. Firstly, the method needs to determine the construction targets and knowledge scope of the domain ontology, and input the target files to be processed. Then, it constructs domain ontology in five stages of preprocessing, concept and term extraction, ontology framework establishment, relation extraction, domain ontology representation and evaluation. Next, we will introduce the implementation process of these five stages in detail.

Figure 1. A research framework for domain ontology construction.
2.1. Preprocessing
The preprocessing of the text is mainly to apply Natural Language Processing (NLP) tools such as LTP\cite{16} to parse each sentence in the input text. We use the regular expression method to segment the text according to the sentence separators “。”, “；”, “!”, “?” and “……” that commonly used in Chinese text. After obtaining the individual complete sentences, NLP tool is used to proceed word segmentation, part-of-speech tagging and named entity recognition for these sentences successively.

Since Chinese language expression is coherent and there is no interval between words, only by parsing them can we better identify the terms, relations, attributes and other information contained in sentences. The parsing effect will affect the processing of subsequent stages.

2.2. Concepts and terms extraction
At present, the commonly used concepts and terms extraction methods are mainly divided into three categories, which are methods based on linguistic knowledge, methods based on statistical calculations, and methods combining linguistic knowledge and statistical calculations\cite{9}. The effective and fast linguistic-based method takes into account the semantic environment of the context in which the concepts and terms are located and processes words with small lexical granularity. So we choose this method for term extraction.

At the same time, we also list important concepts and terms in the domain, that is, in the initial stage of constructing domain ontology, enumerate all the concepts and terms in the domain as much as possible to ensure that the coverage of the domain ontology is large enough, and the number of ontologies involved is sufficient to provide a good foundation for hierarchical construction. We summarize it as a domain dictionary for concepts and terms extraction to help extract concepts and terms contained in sentences, so as to reduce the error rate of NLP tools and improve the recognition rate.

2.3. Ontology framework establishment
In the previous step, we only got a vocabulary with no organizational structure. We need to classify a large number of concepts and terms, and then construct hierarchical relationships among classes to classify all words into corresponding categories. Concepts or terms in the same category are more relevant, while the correlation among concepts or terms in different classes is weaker\cite{17}. In order to improve the quality of ontology, it is necessary to determine the importance of each concept and term, select key concepts and terms, delete concepts and terms that are unnecessary or beyond the scope of the domain, and determine the hierarchical relationship between them. For the construction of the ontology framework, a top-down, bottom-up or comprehensive approach is usually adopted. The meta-ontology is the ontology of ontology, which is the highest level abstraction of the concept in the domain\cite{18}. We use the meta-ontology to help design the ontology framework structure, and we need to pay attention to reuse existing ontology to reduce work consumption and realize knowledge sharing and reuse.

For the design of the meta-ontology, the operation steps are as follows:
- Define the classes and hierarchy of classes in the domain ontology.
- Define the attributes of the classes.
- Define attribute values.
- Create instances.

2.4. Relation extraction
We do dependency parsing for each sentence and construct a relational extraction model based on Chinese grammar rules. According to this model, we extract the relations between concepts or terms in sentences, and finally, combine concepts or terms and relations into inter-word relation triples. Since the relation between concepts or terms is mostly a predicate verb structure, and the predicate verb appears behind the adverbial and appears in front of the object and complement, or these components appear at the same time, that is, relational predicate verb depends on the components before and after
Therefore, we analyse the results according to the dependency parsing and abstract the syntax and semantics of the relation. The resulting relational extraction model representation is shown in figure 2.

The dependency parsing of a sentence and matching a part of the relational extraction model or complete matching the relational extraction model all can extract the inter-word relation triples between concept 1/term 1 and concept 2/term 2 and concept 3/term 3 for the domain ontology construction.

2.5. Domain ontology representation and evaluation
The appropriate ontology description language is selected to encode and formalize the domain ontology established by the above operations so that the computer can recognize and process the domain ontology. At present, ontology description languages mainly include Web Ontology Language (OWL), Resource Description Framework (RDF), Extensible Markup Language (XML), etc.[18].

Formal domain ontology needs to be tested and evaluated, but no complete set of evaluation criteria that are appropriate for each method. In general, we examine whether the constructed domain ontology satisfies the initial requirements, whether it achieves the construction goals, whether it conforms to the construction criteria, and whether it has scalability and reusability, and so on.

3. Construction of the domain ontology of food safety events
We have introduced the domain ontology construction method based on relation extraction in detail. In order to test the effectiveness of this method, we designed a construction experiment of the food safety events domain ontology to verify.

First, we selected the “food safety events” as the research domain, and used the web crawler technology to crawl the texts of food safety events related reports on the network as the processing object. In the meanwhile, we also crawl the national standards, departmental standards, and enterprise standard documents containing a large number of concepts and terms in the field of food safety, such as “GB 2760-2014 Standards for Use of Food Additives”, etc. The concepts and terms of contraband products such as microorganisms, additives, agricultural veterinary drugs, heavy metals, pollutants, etc. that cause food safety events are inquired and used as a domain dictionary.

A related report on a food safety event usually contains two types of information, namely, key elements of the event and irrelevant descriptive information. Here, we define key elements including food classification, additives, pesticide residues, veterinary drug residues, pollutants, food enterprises, people involved, time of occurrence, and location of the nine categories. Through these key elements, we can clearly extract the main information involved in the event. The other information contained in the report, such as product peripheral information and background introduction, have little relevance to the subject of the event and cannot display the specificity of the event. So we define them as irrelevant description information and do not show them in the ontology, thereby improving the quality of the ontology.
As shown in figure 3, we obtained the main categories of domain ontology of food safety events through the above analysis.

![Diagram of food safety event ontology](image)

Figure 3. Main categories of food safety event ontology.

In the food safety event ontology, the hierarchical relations of categories can reach up to six layers, such as the hierarchical structure of the “fruit” category in the food and food category shown in figure 4.
When building the ontology, the open source ontology editing tool protégé\[20\] developed by Stanford University in the United States is generally used, which provides an easy-to-understand graphical interface and convenient extension form. We use the protégé to construct the domain ontology. For the relations obtained by relation extraction, and the inter-word relation triples that combine these relations with concepts or terms, we add them to the corresponding position and finally, we can get the food safety events domain ontology.

For example, in order to clearly observe the hierarchical relation between categories and the inter-word relations extracted from sentences, we only show the relations between “Dried fruit class” in the “Fruit” category and the “Preservatives” in the “Additives” category of the constructed domain ontology of food safety events. The relation triple extracted from the sentence “Preservatives are usually added to dried fruit foods” are (dried fruit foods, added, preservatives), so we can get the relation between “Dried fruit class” and “Preservatives” is “added”. The dotted line shows the “add” relation in figure 5.

Figure 4. The hierarchical structure of the “fruit” category.
We evaluate the ontology from the functional level of ontology. The evaluation indicators generally include consistency and universality, user satisfaction, task consistency, topic degree, and reuse degree. The food safety events ontology is manually defined according to the construction needs and relevant national standard documents and meets each indicator. It is a standard and consistent domain ontology.

4. Conclusion
As a conceptual model modeling tool that can describe information systems at the level of semantic and knowledge, ontology has attracted more and more attention. The construction of domain ontology is a creative and extremely challenging process. This paper proposes a novel method of domain ontology construction based on relation extraction, which processes Chinese text by using the relation extraction method at the semantic level, and constructs domain ontology by combining preprocessing, concepts and terms extraction, domain ontology framework construction and other operations. The construction results applied in the field of food safety events also meet the standards of construction and have good logicality and operability, which provide the basis for the application of domain ontology in the semantic layer. In the future, we will strive to improve the accuracy of the ontology, and gradually realize full automatic construction.

Acknowledgments
This work was supported by the National Key Research and Development Program of China (Project No. 2017YFC1601800).

References
[1] Vasilateanu, A., Goga, N., Tanase, E.A., Marin, I. (2016) Enterprise domain ontology learning from web-based corpus. In: International Conference on Computing, Communication and Networking Technologies, IEEE. Denton TX. pp. 1-6.
[2] Zhang, Z.G. (2008) The study and application of the domain ontology construction method. Doctoral dissertation. Dalian. pp. 1-60.
[3] Shih, C.W., Chen, M.Y., Chu, H.C., Chen, Y.M. (2011) Enhancement of domain ontology construction using a crystallizing approach. Expert Systems with Applications, 38(6): 7544-7557.
[4] Yu, X., Shen, G. (2013) Research on Semi-Automatic Domain Ontology Construction Framework Based on Web Crawler. In: Proceedings of the International Conference on Computer, Networks and Communication Engineering. Beijing. pp. 413-416.

[5] Amith, M.F., He, Z., Bian, J., Lossio-Venture, J.A., Tao, C. (2018) Assessing the Practice of Biomedical Ontology Evaluation: Gaps and Opportunities. Journal of Biomedical Informatics, 80: 1-13.

[6] Tang, A.M., Zhen, Z., Fan, J. (2005) Research on domain ontology construction based on thesaurus. Modern library and information technology, 21(4): 1-5.

[7] Wang, j.D., Wu, P., Zhu, Y.H. (2016) Research on reasoning and construing of movie material domain ontology based on Jena. Computer technology and development, 26(8): 30-34.

[8] Fu, L. (2017) Research on domain ontology construction framework under big data environment. 9 Library, 2017(11): 66-71.

[9] Zhai, Y.J., Wang, F. (2015) Research on ontology construction method of Chinese domain based on text mining. Information science, 2015(6): 3-10.

[10] Du, W.H. (2005) Comparative study on ontology construction methods. Journal of intelligence, 24(10): 24-25.

[11] Yang, Q.F. (2002) Review of Ontology methodology. Computer application research, 19(4): 5-7.

[12] Menzel, C.P., Mayer, R.J. (1990) IDEF5 Ontology Description Capture Method: Concept Paper.

[13] Uschold, M., Gruninger, M. (1996) Ontologies: Principles, methods and applications. The Knowledge Engineering Review, 11(2): 93-136.

[14] Siou, K. (2004) Informational and Computational Equivalence in Comparing Information Modeling Methods. Journal of Database Management (JDM), 15: 73-86.

[15] Li, J., Su, X.L., Qian, P. (2003) Method of constructing domain ontology. Agricultural network information, 2003(7): 7-10.

[16] Che, W.X., Li, Z.H., Liu, T. (2010) LTP: A Chinese Language Technology Platform. In: Proceedings of the Coling 2010: Demonstrations. Beijing, China. pp. 13-16.

[17] Yu, J.D., Li, X.Y., Fan, X.Z. (2008) Design and implementation of domain ontology in information extraction. Journal of university of electronic science and technology, 37(5): 746-749.

[18] Yi, L.T., Zhou, S.Q., Ding, C.S. (2011) Research on modeling method of domain ontology in information extraction. Computer technology and development, 21(10): 23-27.

[19] Li, M.Y., Yang, J. (2016) Open Chinese Entity Relation Extraction Method Based on Dependency Parsing. Computer Engineering, 42(6): 201-207.

[20] Knublauch, H., Fergerson, R.W., Noy, N.F. (2004) The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications. In: Third International Semantic Web Conference. DBLP. Japan. pp. 7-11.