Research on Semi-Automatic Domain Ontology Construction Framework Based on Web Crawler

YU Xue-jun, SHEN Guang-peng
Software Institute of Beijing University of Technology, Beijing, China
yuxuejun@bjut.edu.cn, dapengzhanchi0070@163.com

Abstract - Now the ontology construction is mainly based on manual mode, the whole process requires a lot of manpower and material resources.In this paper we proposed a semi-automatic domain ontology construction framework based on web crawler. The framework can fetch domain data on network and extract semantic knowledge through language methodology and statistical methods. Finally, we construct ontology using the domain ontology modeling method based on extensions. The framework can save the investment of manpower and human resources in the manual construction mode.

Index Terms - Ontology Learning; Web crawler; Semi-automatic construction

1. Introduction

The network is a huge storehouse of knowledge, how to use data effectively from the Internet has become a hot research today. So Semantic Web and Linked Data based on Semantic Web came into being in recent years [1]. All these technologies are based on ontology. However, the domain ontology mainly relies on manual construction. It costs a lot of manpower and material resources. This paper put forward a new fast semi-automatic domain ontology construction framework.

2. The current situation of ontology construction

The methods of ontology construction can be summed up in three types: the manual construction mode、the existing ontology reusing mode and the automatic construction mode. The properties of these three methods are shown in table 1.

TABLE 1 the properties of three methods

Ontology construction mode	properties
The manual construction mode	It started earlier and is the most popular way now. There are also relevant tools. And for different domain there is different construction method. But it requires domain experts to participate in the entire process. So it costs a lot of manpower and material resources.
The existing ontology reusing mode	It is based on existing anthologies. So it can reduce the workload. But in the beginning of the semantic web, there is little ontology that can be reused. Even if there is some ontology that can be reused, it is only part of their content. So it needs some complex work such as ontology mapping, ontology cropping and so on.
The automatic construction mode	It can obtain the ontology knowledge automatically though multidisciplinary technical. It can accelerate the ontology construction, and reduce the human and material resources. However, this technology started late. The researches on it are mainly experimental projects. And it is hard to be full automation in the short term.

Among these methods, the manual construction mode is the most mature way. For different domain there is different construction method, such as the Skeletal Methodology [2] used to build Enterprise Ontology, the METHONTOLOGY used to build Chemical Ontology. And in recent years, many ontology building tools appeared, such as Protégé, Onto Edit, Web ODE and so on [3]. They can provide convenience for manual ontology construction. But this mode relies on domain experts during the whole process. So it costs a lot of human and material resources. The ontology reusing mode is based on the existing ontology and can reduce the workload. But there are also many negative factors. On the one hand, the reusable ontology is not that much. On another, the standards of these ontology is different. So we should pay a lot of complex work on it, such as ontology mapping and ontology cropping. Therefore, the automatic construction mode has become a hot research spot. But scene the ontology is very complex, it is hard to be full automation in the short term. In some phases of the process human intervention is required, so semiautomatic ontology construction appears and is placed high expectations.

Ontology learning is the key part of automatic or semiautomatic ontology construction. The content of ontology learning can be expressed by the formula: O={C, R, Hc, Rel, Ao} [4], C and R represent the collection of concepts and relations; Hc and Rel represent classified and non-classified relations; Ao represents ontology axioms [5]. Now ontology learning mainly focuses on the acquisition of the concepts and their relationships. The ontology learning is mainly based on natural language processing methods, statistical methods and data mining techniques [6]. In foreign countries, this technology started earlier and already has some ontology learning tools, such as Text2Onto, OntoLearn [7]. However in domestic, it started later. Due to the complexity and flexibility of the Chinese language, Chinese ontology learning is more difficult. So the researches in domestic are still in experimental state. Article [8] proposed a method to acquire the archaeological concepts of field and article [9] proposed a pattern-based method to acquire hyponymy concepts.

3. The Semi-Automatic Domain Ontology Framework Based on Web Crawler

Domain ontology is a conceptualized and formalized specification of domain knowledge. The ontology layer is located at the core position in the seven layer model of semantic web [10]. And it laid the foundation for the semantic description of the network resources.

Now a research on semantic integration of domain data based on linked data is in progress in our laboratory. The foundation of the work is to construct a domain ontology which is well defined and has a strong expressive power. Therefore, a domain ontology which can be put into use must have complete domain concepts and semantic revealing ability. So, we should acquire domain concepts and relationships as much as possible in the process of ontology construction.

In this framework, we use the Internet as the data source of domain knowledge. This framework can be divided into three modules: domain corpus acquisition module, semantic knowledge acquisition module, the domain ontology formal representation module. The system framework is shown in Figure 1:

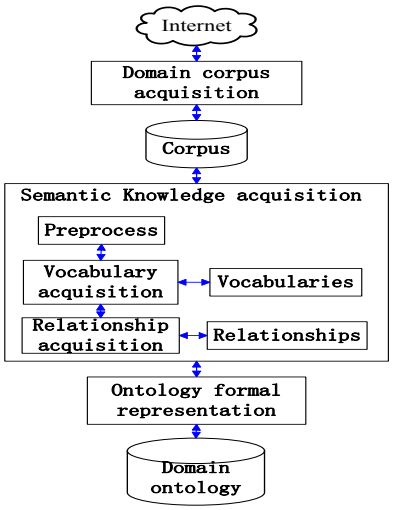

Figure 1 the System Framework of Semi-Automatic Domain Ontology Construction

In this framework, we will extract domain semantic knowledge from original domain data and finally construct a complete ontology. First, the domain corpus acquisition module is designed to obtain rich domain data from the domain website on the Internet. These data is the foundation of all the subsequent work. In this module we use web crawler technology and vertical search technology to ensure the accuracy of the crawling data. Then, the semantic knowledge acquisition module is used to extract domain vocabulary and their relationship from the crawling data. In this module we combine the linguistics technology and ontology learning technology. And we mainly studied the acquisition of domain vocabulary. In this part, we improve the extraction effect of the traditional vocabulary acquisition method by adding another layer of filter, and we will talk about this in chapter 3.2.1. Finally, in the domain ontology formal representation module, we mainly accomplish two things: establish a full ontology system and express the ontology system using ontology language. In the first part, we propose a domain ontology modeling method based on extensions, for example, we can abstract higher level or lower level concept according to the vocabulary we have already acquired. In this way, we can establish a complete ontology system. In the second part, we express the domain ontology in machine-readable form, so the ontology can be easily processed by computers.

4. The Semi-Automatic Domain Ontology Construction Based on Web Crawler

4.1 The acquisition of domain corpus

This module is a base portion of the entire framework. First, we make crawling strategy according to the characteristics of the web pages using vertical search technology. Then, the crawler will obtain domain data starting from seed URLs according to the crawling strategy. This module starts from a general controller which is responsible for receiving seed URLs and scheduling the work of the entire process. Among all the parts of this module, the web data stream processing part is the central component. In this part, we use the popular webpage analytic tool HtmlParser. It resolves the whole web page as a tree structure, and grabs data meeting the requirements using its node filtering function. Finally, it will store the crawling data in text format. In this way, this module improves the accuracy of the captured data and reduces the interference of the useless data. In addition, in order to improve the data crawl rate, this module also uses java multithreading technology to grab data in parallel.

The experiments show that the module has high data crawl rate and data accuracy. Now this module can crawl nearly 300 web pages per minute. The data accuracy is about 80%. And the module has crawled nearly 500M data successfully from the e-commerce website.

4.2 The acquisition of semantic knowledge

4.2.1 The acquisition of domain vocabulary

The domain vocabulary is important to ontology construction. On the one hand, concepts which represent people's recognition of objective things are expressed by domain vocabularies. On the other hand, the relationship between concepts is also expressed by the relationship between vocabularies. Now the methods of vocabulary acquisition can be summarized into two categories: the linguistics-based methods and the statistics-based methods. The linguistics-based methods have a high accuracy but a low recall rate and extensibility. The statistics-based methods have a high extensibility, but their accuracy is lower than the linguistics-based methods. Now we usually use the combination methods of them.

In this framework we preprocess the corpus through linguistics methods, such as word processing using ICTCLAS, POS tagging, stop words removal and so on. Most of the domain vocabularies are complex terms, and they are not registered on word process tools. So the results of the word process tools are usually basic terms. After studying, it is not difficult to find that the domain candidate terms are usually consisting of noun and gerund, for example, complex term "Software Engineer" is composed of "Software" and "Engineer". So we proposed a candidate term extraction

method based on POS Tagging which can extract noun compound phrases. After the experiment, this method has a better result. However, the result also contains some interference vocabulary; some of them are even not phrases, such as “Host Power Quality” and “Computer black screen”. So we will calculate the domain credibility of these terms for further filtering. Now there are many domain credibility calculation methods. We have already realized TF/IDF algorithms and statistical algorithms based on the vocabulary information entropy [11], and we also made corresponding improvement. Here we mainly introduce the TF/IDF algorithm and its improvement. The formula is shown as follows:

$TF.IDF = \sum_{j=1}^{n} \mathrm{tf}_{ij} * \log(\mathrm{n}/\mathrm{df}_{\mathrm{i}})$, tfij represents the number of termi in documentj. dfi represents the number of the documents which contain termi. The formula indicates that the greater the number of termi appears in the document the larger the TF/IDF value is, and termi is more important.

But there is still a problem; it is common that some domain vocabulary appears a very low frequency in a document. So the TF/IDF method is out of use. Taking this into account, we proposed a weight frequency calculation method as a supplementation of TF/IDF algorithms. The weight frequency formula of term a is: $\mathrm{f(a)} = \sum_{\mathrm{j=1}}^{\mathrm{n}} \mathrm{f(a_j)} \Big/ \mathrm{n}$, and $\mathrm{f(a_j)} = \sum_{i=1}^{k} \mathrm{f(a, b_i)} * \mathrm{f(b_i)}/\mathrm{f_m}$,f(aj) represents the weight frequency of the word a in its j-th occurs. bi represents the term that appears along with term a in the sentence. fm represents the maximum frequency of occurrence. f(bi)/fm represents the weight of bi. By this method, though a domain term has a low frequency, as long as it appears along with some high frequency terms, then its weight frequency will increase. That means this term is more likely to be extracted.

In an experiment, “software maintenance” appears only twice in a software engineering document, its TF/IDF value is 0.0313. So it was routed at the end of the result. But through this method, its weight frequency is 0.0951, so this term can be extracted successfully. The weight frequency calculation method as a supplementation of TF/IDF algorithms can improve the experimental results to some extent. We usually use two standards to evaluate the experimental results: Precision and Recall. They are defined as follows:

$\mathrm{Precision} = \mathrm{a}/(\mathrm{a}+\mathrm{b})*100\%$, $\mathrm{Re}\,call = a/(\mathrm{a}+\mathrm{c})*100\%$. a represents the number of the correct words in the result. b represents the number of the wrong words in the result. c represents the number of the correct words that the extraction algorithm cannot recognize. Usually we cannot guarantee these two standards in the same time. If one of them increases, the other usually reduces [12].

Through some experiments, we have got a conclusion that the weight frequency calculation method can significantly improve the recall rate. The experimental result is shown in table 2:

TABLE 2 experimental results

	before filtering	after filtering
Precision	72%	70%
Recall	67%	75%

4.2.2 The acquisition of domain relations

At the present, there are three main methods to get the relationship between concepts: the language model-based methods, dictionary-based methods and statistical-based methods. The language model-based method first need to learn the language mode from a large number of corpus, the famous of them like the Hearst mode [13], which obtains the hyponymy lexical relations. The advantage of this method is its high accuracy, but due to the low frequency concurrency of the language mode, the recall rate is very low, and these rules have strong dominical and low scalability. Dictionary-based method is mainly depend on the currently semantic dictionary that contains a large number of relational terms, such as WordNet、HowNet and so on. This method will be affected by the knowledge cover degree of the semantic dictionary; some vocabulary relation may unable to get in some area. The statistical-based methods getting the relationship mainly depends on co-occurrence frequency, and has become the most popular relationship acquisition method, for example, article[14] proposed a automatically extraction algorithm that based on the semi-structure corpus relationship.

The relationship between concepts is also a key part of domain ontology. Now we have realized the association rules algorithm to get the conceptual relations. This algorithm determines if there is relationship between two terms mainly based on their co-occurrence frequency, for example, the "phone" and "Network standard" have a high frequency of co-occurrence and exceed the threshold, and then we can get this conceptual relation. This algorithm is completely based on statistics. To guarantee the accuracy of the results, in the next step we will take the semantic distance between terms into account.

4.3 The formal representation of the ontology

As is illustrated in chapter 2, this module mainly includes two aspects; we can call them ontology modeling and ontology representation for short. For ontology modeling, there is no uniform standard now. In this framework, the semantic knowledge we have extracted using the former methods generally have domain representative and domain coverage, but they are not a complete system. Given these, we propose a domain ontology modeling method based on extensions according to the semantic knowledge we have extracted. For example, we can extract upper level concept “Computer” from basic concepts “Desktop” and”Laptop”.

In order to allow machines to achieve query and reasoning, we must express the domain ontology in machine-readable form. There are many ontology languages, such as

RDF, OWL and so on. RDF is a veritable ontology language, but it cannot describe complex relationships, such as synonymy relationship and antonymous relationship. OWL is based on RDF, and has a richer vocabulary. So it has better expression ability, and has become a W3C standard. So now people usually use OWL and RDF to express the ontology.

Conclusion

This article proposes a semi-automatic domain ontology construction framework based on web crawler, and describes the process and technologies of various sub-modules. The framework is based on a number of experimental works. We have realized domain corpus crawling method and vocabulary and relationship acquisition algorithms. And finally we have got a good result. In the next step, we will optimize these algorithms to improve performance, and complete the whole process of ontology construction.

References

[1] [1] Tim Berners-Lee, JHendler, OLassila.Semantic Web [J].Scientific American, 2001, 284(5):34-43.

[2] [2] Chen Li. The Overview of Domain Ontology Construction Technology [J].Technology Square, 2011, 06.

[3] [3] Liu Yu_song. Research on Ontology Construction Methods and Development Tools [J].Modern Information, 2009, 29(9):1008-0821.

[4] [4] Yang Fen. Research on concept and relation extraction in ontology learning [D]. Chongqing: Chongqing University, 2010.

[5] [5] Du Xiao_yong, Li Man, Wang Shan. Ontology Learning Research [J]. Journal of Software, 2006,17(9):1837−1847.

[6] [6] Liu Bo_song. Common Ontology Learning Research Based on Web [D]. Hangzhou: Zhejiang University, 2007.

[7] [7] Zhang Nan_nan, Li Guan_yu, Qu Li_ning. The Comparative Analysis of Main Ontology Learning Tools [J]. Micro Computer Information, 2008, 12.

[8] [8] Zhang Chun_xia. Reach on Domain Textual Knowledge Acquisition Method and Its Applications in Archeology [D]. Beijing: Calculation of Chinese academy of sciences, 2005.

[9] [9] Liu Lei, Cao Cun_gen. A Lower Concept Acquisition Method Based on “is a” [J]. Computer Science. 2006.

[10] [10] Gao Zhi_qiang, Pan Yue, Ma Li. Principle and Application of Semantic Web [M]. Beijing: Mechanical industry press, 2009.8.

[11] [11] Sang Ai_ju. Chinese Ontology Learning Techniques Based Text2Onto [D]. Qingdao: Ocean university of China, 2009.

[12] [12] Jin Hai, Yuan Ping_peng. Semantic Web Data Management Technology and Applications [M]. Beijing: Science Press, 2010.2.

[13] [13] Zhang Xi_fu, Dai Yun_hui, Gao Zhi_qiang. Semantic Relationship Extraction from Terminology Dictionary Based on syntactic patterns [J]. Journal of Nanjing normal university, 2008,04.

[14] [14] Qing Hua, Wang Chao_jing, Sun Xia. Semantic Network Automatic Generation Algorithm Based on the Concept of Structured Corpus [J]. Computer Research and Development, 2005,42(3):478-485.