A Design of Knowledge Extraction Process of Work Flow Texts based on Domain Ontology

Yingnan Zhang[1, a]

[1] The Chinese People's Liberation Army Unit 91550, Dalian, 116023, China

[a]email: zhangyingnan_DL@163.com

Keywords: Domain Ontology; Research Text; Knowledge Extraction Process

Abstract. Ontology has the characteristics of consistency, sharing and extensibility, which plays an important role in the study of text knowledge extraction. In the reality of work flow text knowledge extraction, based on the construction of domain ontology, this paper establishes the basic process of work flow text knowledge extraction, analyzes the process of text preprocessing, text instantiation, the problems involved in the three stages of text knowledge reasoning and designs the corresponding solutions for each stage with practical examples in application.

Introduction

Semantics is the meaning of the information symbols exhibited in the computer, including the system meaning and the external meaning [1]. The former refers to the concept system, which involves the environment. Previous research on work flow text knowledge extraction has just conducted at the level of shallow semantic structure preliminarily by using the semantic [2]. In recent years, with the deepening research on ontology, ontology and text knowledge extraction has become the focus of current research [3]. The extraction and application of the knowledge of work flow text emphasize the mining and application of the semantic meaning of the knowledge in the text.

Ontology has the characteristics of consistency, sharing and extensibility, which is important to the study of text knowledge extraction [4]. The constraint to the consistency of the concept is to ensure the independence of knowledge extraction, and clarifies the boundaries between explicit knowledge concept to construct the knowledge architecture of the whole; the concept of ontology sharing is to ensure that the knowledge is able to be shared in heterogeneous platform and remote interaction, asynchronous operation environment; ontology scalable ensures that the text knowledge extraction expands and generates new knowledge. Ontology itself completes semantic structure to ensure the integrity of the knowledge extraction, explicitly explains the logical relationships between these concepts of ontology [5]. Based on description logic or rules of system reasoning mechanism, we manage to ensure the depth accuracy and refine the text knowledge extraction process.

To solve this problem, this paper carries out a ontology technology on the research of knowledge extraction. Based on domain ontology, we establish the basic process of work flow text knowledge extraction, analyze the process text preprocessing, text instantiation, the problems involved in the three stages of text knowledge reasoning and designed the corresponding solutions for each stage with practical examples in application.

A Work Flow Text Knowledge Extraction Process

In the text, the knowledge is essentially based on the vocabulary as the basic storage unit, and the knowledge extraction and reasoning of the work flow text must be carried out around the processing of vocabulary. Based on the above analysis, we design the overall framework of scientific text knowledge extraction and reasoning: first of all to research text pre-process to obtain the intermediate XML text; and then construct semantic sequence extraction model, under the support of domain ontology to text for knowledge extraction and formalization representation; finally,

according to the design idea of ontology reasoning mechanism, design logical inference rules of instantiated knowledge and inference realization. Figure 1 describes the process of knowledge extraction and reasoning for work flow texts.

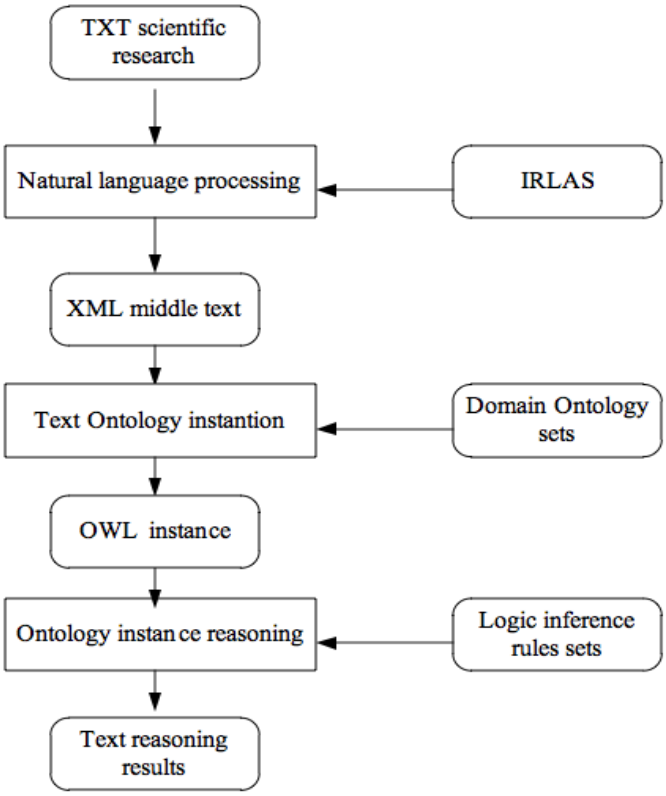

Fig.1. The flow chart of feature-limited text's knowledge extracting and reasoning

Text Preprocessing

The preprocessing of work flow text is mainly focused on the text word segmentation and part of speech tagging. Research text belongs to specific vocabulary domain with professional features, which shows frequently only in its own field and different in plain text word segmentation features. Take equipment index in the field of 'maximum distance' vocabulary for example, it is a full term, but in the plain text it tends to be split as a combination of maximal/distance "two words. Thus it can be seen that it is more suitable to use the method of the complete professional domain dictionary in the field of work flow text. This paper suggests using the IRLAS Chinese lexical analysis system developed by Harbin Institute of Technology LTP Natural Language Processing platform. The system provides a dictionary extension interface that allows a research organization to construct a domain dictionary that satisfies the maximum matching principle in terms of their respective fields. IRLAS tags part of speech by using hidden Markov model, the current relatively mature calculation model. And vocabulary sequence corresponds to a sequence of part of speech. POS probability is calculated to obtain reliable part of speech tagging.

The Basic Process of Text Instantiation

The text instantiation process includes four stages:

(1) The classification of lexical sequence boundaries for each subject in the text;

(2) Semantic acquisition of lexical representation;

(3) The selection of semantic meaning in the semantic sequence;

(4) The extraction and representation of the concept and the associated relation in the semantic sequence. That is the ontology.

The ontology technology is introduced into the text to instantiate each stage, as the semantic retrieval support library, bear the text vocabulary screening, semantic access, support the semantic

concept ontology instantiation, consistent verification, knowledge structure constraints and semantic constraint logic function, and in the meanwhile build a scientific text semantic sequence extraction model. By using the model matching method to extract the text of the relating concepts and concept relationships between the specific content, we achieve the instantiation of the corresponding concepts in the ontology, and ultimately complete the task of formal representation of research text knowledge extraction.

Text Knowledge Reasoning

It is an important method to get the knowledge reuse and further mime knowledge on the association and constrains between text knowledge, and an important prerequisite for information associative retrieval, automatic summarization, machine learning and other fields, as well as has the important applications in the application of knowledge. Machine reasoning about the knowledge of work flow text, in essence, is the reasoning of formal representation of knowledge. This paper views the acquired form of knowledge through the method of knowledge extraction as the object of inference. Through the study and analysis of realization mechanism of reasoning machine, we analyzes the characteristics of three kinds of logical reasoning, extends the types of reasoning, and preliminarily studies the reasoning method based on ontology technology research of text knowledge.

Logical reasoning in the instances of knowledge text can be divided into two parts. The first part is to set up two instances which can access rule of inference. Only when the instances meet the access rules between each other, the inference operation could work. Otherwise there is no necessity and significance in reasoning; the second part is to establish the logical rules to support instance reasoning, according to application needs. In this paper, these two kinds of reasoning are called access reasoning and case reasoning.

Four tuple structures are used to express the knowledge of work flow, and the reasoning of the above reasoning object is extended to the four tuple of reasoning.

(1) Access reasoning

Admission reasoning refers to the two instances sets M_1 and M_2: whether they qualify the reasoning of reasoning to judge. I in the four tuple structure $M = (I, \{P\}, \{P_{\text{condition}}\}, \{M'\})$in the ontology class and $\{P_{\text{condition}}\}$、$\{M'\}$are the objects of access reasoning. The value of these objects in the two examples of centralized access rules is to focus on the corresponding rules of reasoning to get a logical result, as an example set to decide whether the reasoning qualifications could function.

(2) Case reasoning

Case reasoning refers to logic reasoning reached by a set of central inference logic rules which are determined when the $\{P\}$in M in two instances choose the value, so as to produce the logical reasoning results between two examples.

Both access reasoning and case reasoning are related with the construction of the inference rule set and the realization of the inference rule algorithm. The construction of rule set will follow the logic rules of mixed chain.

Access inference rules for the object is I in M of the four tuple structure of the ontology classes and $\{ P_{\text{condition}} \} M'$, and examples of inference rules point to the $\{P\}$ in M. Practical purposes in the field of application of these objects are designed to meet the purpose of the logical rules algorithm. In order to realize the user interaction of the logical algorithm, this paper constructs the logical inference rule mapping database as the interface between the user and the logic algorithm. Users can modify the object of each rule algorithm to participate in reasoning through the interface, so as to meet the different objects based on different logic rules of reasoning. Take equipment research indicators in the field as an example, we build the mapping part of the contents of the library as shown in Table 1.

The reasoning process framework of the whole set of examples is shown in Figure 2.

Tab.1. Partial content in the address library about logic reasoning rules

Logical rules tag	instance attribute A	attribute a	instance attribute B	attribute b
Low	Max detection tange	Numeric data	Max detection tange	Max numeric data
high	Max detection tange	Numeric data	Max detection tange	Min numeric data
Equal	Max detection tange	Numeric data	Max detection tange	Numeric data

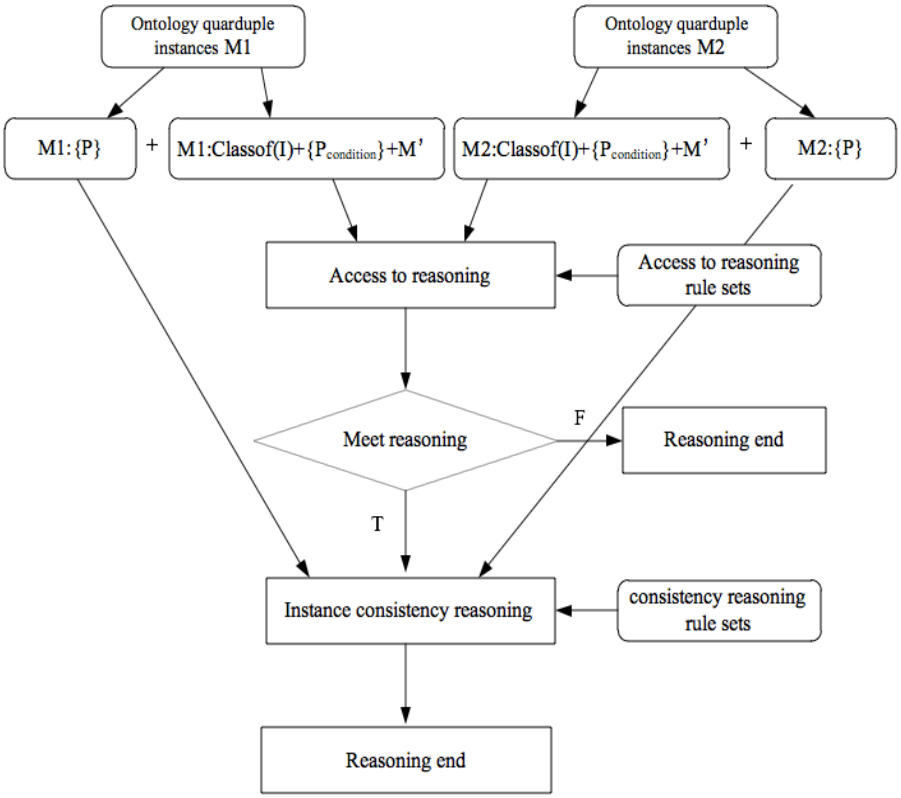

Fig.2. The flow chart of quarduple instances collection's reasoning

Conclusion

Previous research on scientific text knowledge extraction has just been kept at the level of shallow semantic structure, by using the semantic on elementary grade. In recent years, with the deepening research on ontology, ontology has its own characteristics: sharing, scalability, which play an important role in the study of text knowledge extraction. The combination of ontology and text knowledge extraction research has become the research hot zone. To solve this problem, this paper carries out an ontology technology on the research of knowledge extraction. Based on domain ontology, we establish the basic process of scientific text knowledge extraction, analyze the process text preprocessing, text instantiation, the problems involved in the three stages of text knowledge reasoning and designed the corresponding solutions for each stage with practical examples in application.

References

[1] Xu Baoxiang. Research on the method of knowledge representation [J]. Information Science, 2007, 25 (5):690-694.

[2] Huang Yinghui, Li Guanyu. Semantic and semantic Web semantic [J]. Computer engineering and application of semantic Web, 2008, 44 (16):23-26.

[3] Xu Dezhi, Wang Zhiyong. A comparative analysis and Research on the main ontology reasoning tools [J]. Modern library and information technology, 2006, 144 (12):12-15, 77.

[4] Li Na, Ren. Switzerland winsome thesaurus, taxonomy and ontology of modern distributed [J]. information, 2007 (12):122-126.

[5] Angie Hou. Research on semi-automatic construction method of domain ontology [J]. Library theory and practice, 2007 (5):26-27, 38.