Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(12):291-296

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

Research on construction of natural language processing system based on semantic web ontology

Yi Wang*, Jianming Zhang and Yan Xu

Department of Information Engineer, Guangdong Polytechnic, Guangdong Foshan, China

ABSTRACT

The semantic web is an extension of World Wide Web and kind of intelligent network, and it can understand human language. As foundation of the semantic web, Ontology is composed by the finite terms and their relationship. Natural language is the most direct method for knowledge representation. Natural language understanding provides a new way to the research for expert system knowledge acquisition. This paper first analyzes the structure and function of the natural language processing system and proposes construction of natural language processing system based on Semantic Web ontology. The experiments prove that the natural language processing system based on ontology is more efficient.

Keywords: Semantic Web; Ontology; RDF; Natural Language Processing.

INTRODUCTION

Natural language is a symbol system extremely complex. Although a person can handle very skillfully in their mother tongue, but not the expression pattern and language of their native law of constitution, meaning the rules to use computer can accept the way thoroughly clear. The traditional linguistics is developed in the absence of computer reference conditions, although for natural language understanding and accumulated valuable wealth, but that is about to the human, really let the linguistic knowledge into computer operation, is not so simple, nor so fuzzy. Machine Translation natural language understands of the earliest research field. The research of theory and technology in the early limitation, technical level of machine translation system developed is low, can not meet the requirements of practical applications [1]. Woods (Woods) LUNAR system, Vernon Gander (Winogand) of the SHRDLU system and the shank (Schank) of the MARGIE system is a typical example of language understanding dialogue system.

After entering 80 age, the application of natural language understanding extensively, machine learning research is very active, and the emergence of many higher level of practical system. The system is one of important research results show that the natural language understanding, natural language understanding has made breakthrough progress in theory and application.

The difference between the Semantic Web Semantic Web and the world wide web is the extension of the world wide web, but the web has the very big difference, mainly displays in: the object oriented L is different: the current web mainly use HTML expression Webpage content. The use of HTML markers Webpage indeed can express some control Webpage display format of such information, so as to make people think that computers really can "understand" our intention. 2 works in a different way, object semantic web and the World Wide Web for different ways of working, their nature is also different. The world wide web is mainly oriented to "people", so most of its work is done by the people, including information collection, retrieval, sorting, sorting and analysis etc.. The semantic information and the semantic web by adding some computers can be "understood", you can free the people from the various types of tedious work, the use of "intelligent agents" to help do most of the work mentioned above.

A typical example is the use of information retrieval, intelligent search agent; the semantic web will give people the information content of real need, unlike today's search engines like output useless tens of thousands of search results. This paper presents research on construction of natural language processing system based on semantic Web Ontology.

1. Design and analysis of semantic web and Ontology

All the ontology is as the foundation of the semantic web. I believe in part because of comprehensive idea of the semantic web, in part because of the semantic web will bring what interests dispute. The semantic web based on the world wide web, the world wide web is all inclusive network, so the semantic web idea is also a contains all the semantic web [2]. The semantic web is expected to bring benefits to the controversy is, once all information clearly and unambiguously be unified semantic annotation, semantic web can quickly search and reasoning, thus eliminating the current web confusion caused confusion and contradiction. This annotation ontology classification based on Semantic Web, once to a certain "quality", it would be very effective.

We use the world wide web, is actually a storage and sharing images, text media, computer can only see a bunch of text or image, its content can not be identified. The World Wide Web information, if you want the computer to deal with it, we must first of all the original information of these information processing into computer understandable only after the treatment, this is very troublesome thing. And the establishment of the semantic web will things becomes much simpler.

The semantic web is the essence of web changes, mainly the development of its task is to make the data more convenient for computer processing and searching. The ultimate goal is to let the user into God Almighty, the massive resources on the Internet to reach almost omniscient degree, computer can find the information you need in these resources, thus in World Wide Web existing information isolated island, developed into a giant database.

The use of text is processing technology, how to improve the situation? One solution is to use existing expression to represent the content on the web, and development based on artificial intelligence and computational linguistics is becoming more and more complicated techniques to solve the existing problems [3]. This approach has so far been explored much, despite some progress, but the task is still too difficult, as is shown by equation (1).

$$R = [r_{ij}]_p xp = \frac{Z^T Z}{n-1} \tag{1}$$

Another way is to use a more easily processed by machines out said method to describe the content on the web, and that the technology to use this representation convenience. We call this revolutionary program known as the semantic web movement. Generally speaking, ontology forms to describe a domain (domain of discourse). The composition of a typical body by the relationship is between the finite terms and their. The term (term) refers to the final conclusion in the domain of the important concepts (object type). For example, to a university as the domain, then the staff, students, curriculum, classroom and is the subject of some important concepts.

The relationships between concepts usually include a class hierarchy. A hierarchical structure for prescribed C is another kind of C/ subclass, if all the objects in C are included in the C/. For example, all faculties are faculty, as is shown in Figure 1.

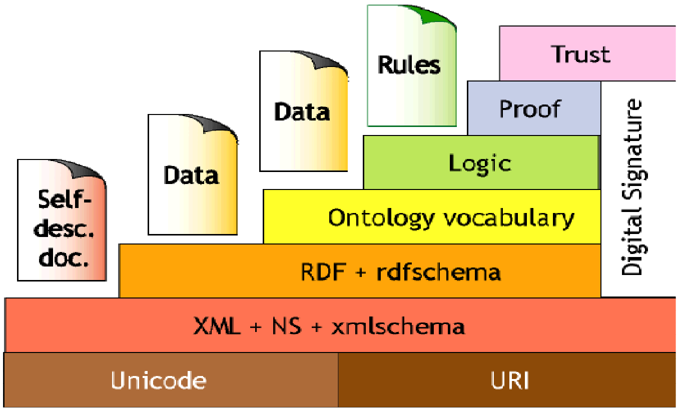

Fig. 1. System structure diagram of the Semantic Web

The realization of the semantic web is mainly related to the core issue of the following: resource description, which actually consists of three small problems, first of all is how to encode the resources [4]. We know that the network English characters except outside, also include Chinese, French, Russian, Han Wen and other various character sets, inside the computer, must be between the coding can be convenient computer unified sharing and exchange of these different character set. The second is how to the resource location, so that a resource is different from other resources specific URL. The third is to describe the resource itself and the represented by what way and means of meaning.

The semantic description of resources, the semantic description of purpose is to understand and deal with convenient computer "". For example, how can we ensure that the text "computer" is the expression of an entity concept rather than a string of numbers? How to make the computer understand the text "Mao Zedong" refers to a person rather than a machine.

Some people are committed to study to indicate the likelihood of RDF data. However, it is the final conclusion of these studies and to fully consider the scale problem. One of the study's proposed RDF data storage based on path, but this method based on path eventually in the relational database based on: it is the sub graph stored in different relational tables. Therefore, such systems cannot provide massive RDF data query scale. Others focus on the measure of semantic similarity network internal and estimation method is used to selectively to optimize the RDF data query; these methods based on memory map implementation, there are still limitations on the scope of it.

As the language definition is currently a very active research field in network based knowledge representation, there are many proposals and new standards. The most important of which are RDF mode and DAML+OIL (recently re defined as OWL), the latter defined on the former one. In addition, there are XML model and Topic Maps is sometimes seen as a knowledge representation language.

In short, the semantic web is a more colorful, more personalized network, you can give it a high degree of trust, let it help you filter out what you don't like the content, so the network is more like your own network.

2. Construction of natural language understanding and processing systems

Language is with the word as the basic unit of grammar, vocabulary and by the dominant can constitute meaningful and understandable sentences, sentence according to certain forms of re form text. Vocabulary can be divided into words and idioms. The idiom is the fixed combination of some words, such as the Chinese idiom. Word is composed of morphemes, "teacher" is composed of "teaching" and "teacher" this two morphemes. Morpheme is the smallest meaningful form word units. "Teach" the morpheme itself has the education and guidance meaning, "normal" contains the meaning of "man".

Grammar is the language of the organization. The rules of grammar restricts how the morphemes words, words form phrases and sentences. Language is formed in the restrict relationship in this tight [5]. Use of morphemes word rules called word formation rules. Whenever you use the interpretation of the model to explain relationships or action, these interpretation model are basically in support of a tree and (AND) tree, the tree and the tree reduces is explaining the relationship or action and the difference between the known relationship and action. And the root node of the tree corresponds to the relationship or action is explained; leaf node corresponds to a causal interpretation model in relation to the connected chain, as is shown by equation (2).

$$I(t) = 1 - R_0 - S + \frac{1}{\sigma}\ln(\frac{S}{S_0}) \tag{2}$$

If this is a good exercise, so the tree with the general should be a good bridge between them. Therefore, MACBETH uses a new description and the tree in order to store for future use. Practice mentioned precedents explain part of the template generation reunion, because the new description can include many model, so it is called Reunion (recollection), which means to use the new method to collect the knowledge, or knowledge is to collect. Using Macbeth and Greed to practice as a simple example of precedent, and it is precedent and practice by the reunion.

For the grammar G = (VT, VN, S, R), if S=*=> is called alpha, alpha is a sentence. Containing only terminator good sentence is a sentence. Grammar G generated sentences all is a language, it will be denoted as L (G), as in equation (3) [6].

$$L(G)=\{\alpha|S=^{+}=>\alpha \;\&\alpha\in V_T^{*}\} \tag{3}$$

For the grammar G1, G2, if L (G1) = L (G2), called G1 and G2 grammar is equivalent to it.

Augmented transition network ATN by Woods (Woods) proposed in 1970. ATN is composed of a group of network; each network has a network, each arc on the conditions for extended operating conditions with. The conditions and operation with register methods to achieve on the register in each component structure analysis of tree, used to store the syntactic features and syntactic features, and the operating conditions which will be constantly access and set.

Algorithm
For a sentence S, there are many types of participle, however the final output results can have only one, so we choose from which an output. Of course, it is the choice of a maximal probability. For example: S: differences of opinion

Word segmentation method W1: intentional / see / divergence
Word segmentation method W2: have / opinion / divergence
Solving methods as is shown by equation (4).

$$MAX(P(W1 \mid S), P(W2 \mid S))$$
$$P(W \mid S) = \frac{P(S \mid W)P(W)}{P(S)} \approx P(W) \tag{4}$$

The P (WN) for the N word appears in the corpus frequency. Natural language is a tool for people to exchange ideas. Since communication is the thought that thought itself in the computer organization structure becomes more important. In artificial intelligence, this is “knowledge representation” of the problem. Can say, said the breakthrough on the problem in the knowledge, will have a decisive influence on the progress of natural language understanding.

Study on the application of natural language understanding to carry out extensive, machine learning research is very active, and the emergence of many higher level of practical system. Among the more famous are American METAL and LOGOS, PIVOT of Japan and HICAT, France's ARIANE and Germany's SUSY system; the system is one of important research results show that the natural language understanding, natural language understanding has made breakthrough progress in theory and application. Intelligent computer study proposed since 80 and, also puts forward new requirements on natural language understanding. In recent years, and puts forward the research of multimedia computer.

The matching system work by reverse is chain inference rules based on the idea of it. Two key ideas that state reduction and backward chaining inference are matching: (a) the precedent as referred to in the preceding paragraph consequent rules to deal with - source. (b) The practice as the assertions database (database of assertions).

To see these two ideas have any help, let us first consider Macbeth as mentioned in the preceding paragraph precedent - consequent rule source. Check that, there are 5 chains is explained one or more causal chain, the chain of causation of the 5 chain is connected with the other chain up. Each of these 5 is explained in the chain, can be easily converted to the preceding rules - back.

Analysis of the sentence procedure is as follows: (1) using LFG syntax analysis to obtain C-structure with context free grammar, do not consider the subscript grammar; the C-structure is a direct component tree; (2) each non leaf node is defined as a variable, according to the subscript lexical rules and grammar rules in the establishment, function description (equations) [7].

Type 3 grammars or regular grammar (RG) another definition is: let G be a type 0 grammar, if produce each type of G A B alpha or A 61664 alpha, alpha epsilon VT*, A, B in VN, then the grammar G is type 3 grammar or regular grammar (RG) or left linear grammar, as is shown by equation(5).

$$K_k = P(k \mid k-1)H^T (HP(k \mid k-1)H^T + R)^{-1} \tag{5}$$

Great changes have taken place in the field of natural language processing. Two obvious characteristics of this kind of change is: (1) to the system input and the requirements of natural language processing system developed can handle large-scale real texts, but not as the previous research system that can handle only a few entries and typical sentence. Only in this way, the developed system has real practical value.

(2) the output of the system, whereas the real understanding of natural language is very difficult, the system does not require deep understanding of natural language text, but to be able to extract useful information. For example, automatic extraction of index terms, filtering, retrieval of natural language text, automatic extraction of important information, automatic abstract etc..

3. Construction of natural language processing system based on semantic Web Ontology

For human use of language is still the most natural representation of knowledge, which requires the fuzzy media into structured knowledge, access by other network service subject and the semantic web. Therefore, language technology tools are very important in the semantic web development, including the following three areas: knowledge markup and ontology knowledge development, intelligent interface.

Knowledge markup: turning the Internet into semantic web imply that ontology knowledge markup for extensive comments on document based on. These documents are mostly composed of free text in different languages, only use the language of technology tools automatically effectively marked. Ontology knowledge development: the ontological knowledge with time in different applications and rapid development. Therefore, learning theory combined with natural language processing semi automatic ontology (text mining, information extraction) and machine learning, is central to effective application.

Therefore, the analysis and understanding of the process of language should also be a hierarchical process. Many modern linguists put this process is divided into 5 levels: phonological analysis, lexical analysis, syntax analysis and semantic analysis and pragmatic analysis. Although this is not between the layers are completely isolated, but this division of hierarchical do indeed help to better reflect the composition of the language itself.

(1) the availability of content, namely semantic web pages and is constructed based on the Ontology is still very little; (2) the development and evolution of ontology, including methods for development, the development process of the core ontology in all fields and technical support, ontology evolution and tagging and version control problem; (3) contents scalability, namely the semantic web content, how in a scalable way to manage it, including how to organize, store and search etc..

Generally speaking, has two branch coefficient and the depth of the decision tree D containing 2D leaves. Therefore, if you want to identify several objects, then d must be large enough to ensure that 2D is more than or equal to n. On both sides of the logarithm shows that the required number of comparison (corresponding to the depth of the tree is log2n D). If 8 objects, comparing the number of log2 (23) =3 instead of (8-1) =7, then save is not significant, as is shown by equation (6).

$$(1-B)^{\delta} = \sum_{k=0}^{\infty} \binom{\delta}{k} (-1)^k B^k \qquad (6)$$

In XML, we have a data model, finally, it on the structure of data, text structure data, any interaction between the semi structure data, can be highly addressing. The real convergence factors behind the trend are getting better and better business intelligence, through its advantages [8]. When the unstructured information is to a managed resource, and it is coordinated for the day of the secret program and it is such as search and obedience.

A formal type Sigma on the conversion of statute law G = (VT, VN, S, R). To include VT = sigma, determine the following measures by the production and the elements of the VN: a) for any formal type R, select a nonterminal S generation S 61664 R, and S as the identification symbol G. B) if x and y are regular type, to shape such as production, A XY heavy 61664 written: A xB, B y two generation 61664 type, where B is a non terminating new selection operators, namely the B in VN.

The semantic web architecture consists of seven layers, from bottom to top are encoded positioning layer (Unicode + URI), XML structure layer (XML + NS + XMLSchema), resource description layer (RDF + rdfschema) (Ontology vocabulary), the ontology layer, logic layer (Logic), proof layer (Proof) and trust layer (Trust). The mutual relation between each layer and it is layer by layer from bottom to top to expand to form a function gradually enhanced system.

A valid XML document in the form of guarantee good at the same time, the value of the elements, and the frequency of the occurrence of nested arrangement or order must be consistent with the elements in the DTD type declaration. With the entity declaration is similar; DTD also can be divided into the internal DTD and external DTD. If the DTD and the XML document in a document, the DTD is called the internal DTD. RDF is a description of objects ("resources") and object relationship data model, and provides a simple semantics for this kind of data model, the data model can be used to represent the XML grammar. RDF Schema is a description language attribute characterizations of RDF resource and class vocabulary, semantic hierarchy with a general to specific about the relations between these properties and classes of the. Finally, through the experimental comparison based on the Semantic Web Ontology DTD and RDF to build a natural language processing system, the experimental structure as shown in figure 2.

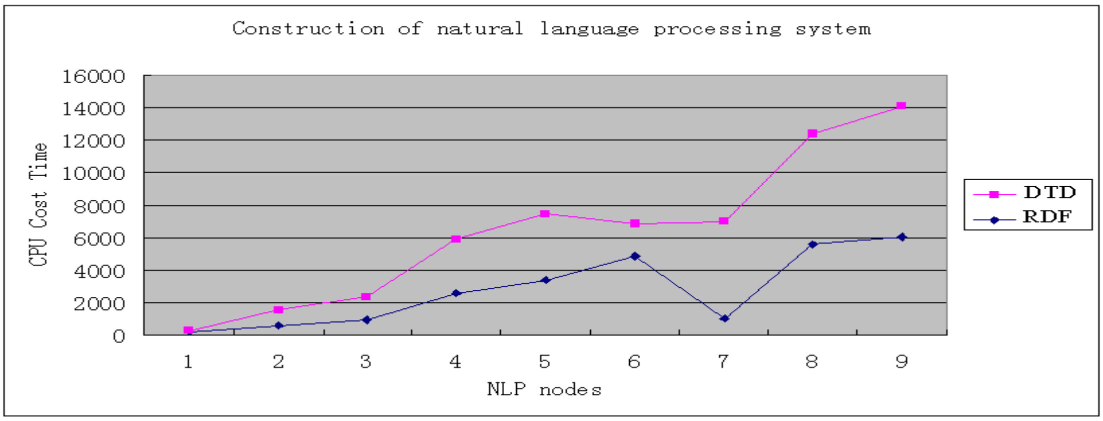

Fig. 2. Comparison results of construction of natural language processing system based on semantic Web Ontology DTD with RDF

The table above RDF data class relationship structure imposed on semi structured. These are not imposed structure exist naturally leads to the expression of sparse many NULL values in the table. The syntactic analysis of the linear order between words will transform into a display structure of words related to other words linked. Semantic analysis of various significances is attributed to structural analysis program established by syntactic, namely, the mapping transformation between syntactic structure and task in the field of object.

CONCLUSION

In a word, semantic web research trying to relate information access problems: establishing the system to help the user location, proofreading, comparison, the control contents. Similarly, we believe that the user should be able to use everyday language to access information, the semantic web concept to linguistics excitation based. Natural language is the rapid adoption of intuitive, easy to use, does not need special training. This paper presents research on the construction of natural language semantic Web Ontology processing system based on. In Natural Language Processing significance: on the one hand, if the computer can understand, natural language processing, will be a major breakthrough in computer technology; on the other hand, Natural Language Processing help unravel the mysteries of human highly intelligent, deepen our understanding of the nature of language ability and thinking.

Acknowledgements
This paper is supported by Training plan of Guangdong Province outstanding young teachers in Higher Education Institutions (Yq2013172), and also is supported by College Level Research Project of Guangdong Polytechnic (K201302).

REFERENCES

[1]. LI Ning; XU Shoukun; LI Bo; Shi Lin. *AISS*, **2012**, 4(1), 154 - 161.
[2]. Tao He; Liping Li; Huazhong Li; Jiangang Chen. *IJACT*, **2012**, 4(5), 24 - 31.
[3]. Wang Qifeng; Zhou Liandong; Lv Hongbo. *JDCTA*, **2011**, 5(10), 127 - 135.
[4]. Feng-Jing Shao; Shun-Yao Wu; Jin-Long Wang; Chun-Yuan Tian. *JCIT*, **2013**, 8(1), 765 - 771.
[5]. Pei Yin; Hongwei Wang; Wei Wang. *AISS*, **2012**, 4(15), 33 - 41.
[6]. Zhijuan Deng; Shaojun Zhong. *IJACT*, **2013**, 5(1), 668 - 677.
[7]. Nurfadhlina Mohd Sharef; Shahrul Azman Noah. *JDCTA*, **2013**, 7(13), 53 - 63.
[8]. Rujuan Wang; Qiushuang Wang; Shuai Lu; Yu Liu; Lei Liu. *IJACT*, **2012**,4(10), 185 - 194.