

Analysis of natural language understanding technology based on Semantic Web ontology

Yi Wang

Department of Information Engineer
Guangdong Polytechnic
Foshan, China
wangyifsfz@163.com

Jianming Zhang

Department of Information Engineer
Guangdong Polytechnic
Foshan, China

Zhenjie Cao

International School
Beijing University of Posts and Telecommunications
Beijing, China
470413717@qq.com

Abstract—The key technology of the semantic web are include: ontology, metadata including logic and reasoning and intelligent agents display. Ontology is a formal, explicit specification of conceptualization of the domain knowledge. This paper analyses problems of natural language understanding including: the expression of the target representation complexity, mapping type diversity, the difference of every element in the source of interaction between the degrees of it. This paper presents analysis of natural language understanding technology based on Semantic Web ontology. The experiment result demonstrates that treatment effect of the semantic ontology can improve the natural language understanding.

Keywords- *Semantic Web; Natural language understanding; Ontology; RDF; XML*

I. INTRODUCTION

Natural language understanding is a fascinating in computer science, challenging. From the computer science, especially from the viewpoint of artificial intelligence, natural language understanding task is to build a computer model; the computer model can give as humans do, analysis of natural language understanding and answer (i.e. the various people's daily use of colloquial language) results.

Language understanding is defined from natural language into a machine (computer system) a mapping between internal. From the macro perspective, language understanding is refers to the machine can perform certain language features desired by the human [1]. The understanding of natural language is a very difficult task. Natural language is not only the semantic, grammar and speech problems, and there are also some ambiguous problems. Specifically, natural language understanding difficulty is caused by the following 3 factors: the complexity of object representation; diversity mapping type; the difference of each element in the expression of the degree of interaction between the source..

The semantic web is the prospect of hope so that everyone can be like the release Webpage that facilitate the release of interconnected machine process able information. The basis is called description standard logic data framework model for resource. Now most of the content on the Internet is designed for humans to read, not for computer programs automatically processed by Webpage content. The semantic web is to extend the current WWW, which can express can be a machine (computer) meaning "understanding", in order to make it easy for people and machine (computer) and the interaction and cooperation between machine and machine. In the Semantic Web "semantics", in fact refers to the meaning of a text.

All the data information of grid resources include organization and are required to make the data meaningful media data. These data may be structured or unstructured, can be stored in any location, such as a database, a local file system, or mail server, or created by any application. This vision is based on semantic grid of visual information grid.

However, the understanding of natural language is a very difficult task [2]. The computer system is also very not easy. This is the middle of a very complex encoding and decoding problem for a lot of. A process that can understand natural language computer system just like a person, and it is need context knowledge and reasoning according to these knowledge and information.

In the environment of the World Wide Web, ontology provides a common understanding of a given domain. This consensus is necessary to eliminate the term difference. For example, the postal code in one application may be just the area code in another application. Another case is two different applications may use the same term to represent different meanings. This paper presents a natural language understanding technology of Semantic Web Ontology Based analysis.

II. RESEARCH ON APPLICATION OF ONTOLOGY AND RELATED TECHNOLOGIES IN THE SEMANTIC WEB

The semantic web is the ideas of the future of the World Wide Web, in which all network based knowledge is clearly, formal code, to allow for more and more intelligent and autonomous subject. Based on the ontology definition network to represent the knowledge, can be and the main body, and the use of the ontology knowledge documents and databases (text, multimedia) marking the understanding of knowledge. Mark describes the current knowledge in the so-called metadata defined in the collection and use of limited application.

RDF three elements have been stored in a three tuple in a large table, which brings the serious problem of scale. Some other methods by trying to kind of relational tables to alleviate these problems, these kind of relational tables will be many three tuple set of features as its property. However, these methods still can not overcome the disadvantages of scale. Recent efforts to solve this problem is to focus on overcoming the problem scale, produced the appearance of vertical partitioning method. In this method a three element (a relational table with three columns, each column represents a resource) are written to the N two rows of the table, where N is the number of independent characteristic data.

The semantic web is to be on the World Wide Web on adding some semantic information can be computer "understanding", it is in convenient for people to read and use at the same time, also facilitate exchanges and cooperation between the computers [3]. Therefore, the World Wide Web is the main object oriented "person", and the object oriented semantic network is mainly "machine". The organization of information in different ways, due to two different objects oriented, therefore in the way of information organization will naturally have very big difference. The World Wide Web in the organization of information resource is mainly centered on "people", according to people's thinking habit and convenience of network information resources organization, as is shown by equation(1). The Semantic Web Information Resources in the organization when it must balance the computer to the text content "understand" and their mutual exchange and communication.

$$p' = W(p, A, D) = Ap + D = \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix} p + \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} \quad (1)$$

Where Merge is a merging process, it associates the entire basic element in the parentheses together, that is the same node merge into one, thus forming a semantic network.

At present, the main obstacle to provide greater support is for Internet users, online content is not machine interpretation. Of course, there are some tools to retrieve the document, cut it into smaller parts, check the spelling and word frequency statistics etc.. But, when it comes to explain the sentence meaning and extract useful information for the user, the existing software ability is limited.

The semantic network can describe more complex semantic relations between things; here are several

common aggregation relationships: if the lower concept is one aspect of its upper concept or a part, said they are aggregation relationship. One of the most promising features of the semantic web is to see the concept and connect to your application's domain inference. An information grid contains only one data resource. The ability to connect different resources and the origin of insight is a key part of the process. As an example, we want to get a public personnel survey information and its social network from the investigators (here refers to the co authors) into the list. Drag and exit the new directory to the XDB database, the system was told to the new directory.

The Semantic Web "is different from the existing world wide web, the data is mainly used by humans, a new generation of WWW will provide also can processing for computer data, which will make the intelligent service a large number of possible"; the research activities of the goal of the semantic web is "the development of a series of computer can understand and express semantic information processing language and technology, automatic reasoning to support network environment widely and effectively".

Because the XML does not include any display format information, so the XML document needs the help of stylesheet (Style Sheet) display. This has one advantage, that is, the same XML document can be displayed using different methods. Style sheets generally exists as a separate document form, the description of the structured document expressive modes, including structured document display mode, the printing effect etc.. And describes the document display compared way HTML style sheet has many outstanding advantages, such as the performance effect of diverse, easy to read, small volume, good maintainability etc.. It is because the style sheet can make the display and data of XML phase separation and a variety of advantages style, so W3C recommends using style sheets to describe structured documents show. Semantic, ontology knowledge using XML mode, RDF mode or DAML+OIL are based on the concept of namespaces, which defines any XML, RDF, DAML+OIL expressions to represent the environment, as is shown by Fig .1 [4].

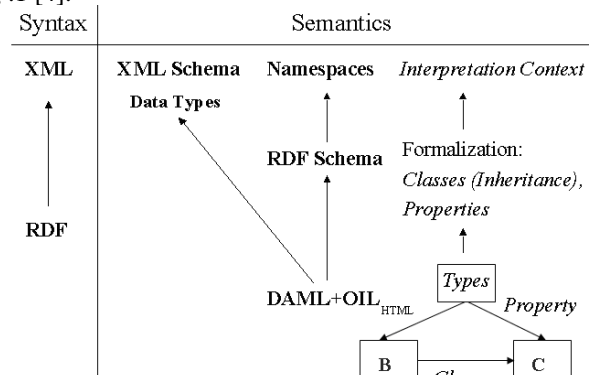


Figure 1. Semantic Web ontology structure graph.

RDF data are decomposed into a number of single states, to directly store into the relationship or hash table. The system of simple state based query can get ideal results. But this lack of a portion of the three tuple or based on is the two part, the result is a set of resources to

improve the missing part, but also can not represent the query RDF data is the most important.

In the architecture of the semantic web, code positioning layer (Unicode + URI) at the bottom, is the foundation of the semantic web, where Unicode is responsible for handling the resource code, URI is responsible for the resource identifier. Only it is based on coding and identification of resources on further processing of resources. The realization of the semantic web will make the cooperation among enterprises do not need too much overhead. Abstract the domain model with the help of the standard can eliminate the terminology differences between implementations use different terms, a document written translation, so as to realize the data exchange. Auction, negotiation and drafting contracts will be software agent automatic (or semi-automatic) to complete it.

III. NATURAL LANGUAGE UNDERSTANDING TECHNOLOGY COMPARATIVE ANALYSIS

The knowledge base is the second factor system design affects learning. The knowledge representation has many forms, such as feature vector, first-order logic statement, production rules, the semantic network and frame etc.. These representations have their own characteristics, in the choice of said 4 aspects consideration to the following way: (1) strong expression ability. (2) Ease of reasoning. (3) Easy to modify the knowledge base. (4) Knowledge representation is easy to expand.

The recognition process to reach the f state (final state), so the sentence was successfully identified. The results of the analysis are shown in equation (2) [5]. From the above process can be seen, this sentence can also go the other arcs in the network, such as the word ducks can also go arc, but the next swallow couldn't find the right arc. Corresponding to this path, the sentence will be rejecting. As a result, shall find out all possible paths in the process of network recognition, so the algorithm to be used in parallel or backtracking mechanism.

$$\phi(t) = \sqrt{2} \sum_i h(k) \phi(2t - i) \quad (2)$$

Assume that U and V are regular type Sigma on regular sets, they said they were recorded as L (U) and L (V), then, (U|V), (U.V) and (U) * are also regular type, regular sets they represent respectively (U, L) L (V), L (U) L (V) (connected set (L) and (U)) * (Bi Bao). The above procedure is used only by the limited time and expressions defined is the formal type Sigma on, said only by these regular expression character set is regular set Sigma on.

On the understanding of the general computer natural language is evaluated from the practical point of view. If the computer has realized the man-machine conversation, or Machine Translation, or automatic abstract language information processing functions, is that the computer has the ability of natural language understanding. There are two aspects existing problems: on the one hand, the grammar so far are limited to the analysis of a sentence in isolation, the system research of context and conversation environment is a lack of restraint and influence of this sentence, so analysis of the ambiguity, ellipsis of words,

pronouns, the same word in different situations and different meanings such problems by different people.

Although the language representation into a series of text symbol or a string of voice flow, but its interior is in fact a hierarchical structure, from the formation of language can clearly see this hierarchy. A literal expression sentence is composed of morphemes and word - or word, phrase or sentence, and the sentence is used to express the voice from the phoneme, syllable to syllable words tone words, and each of these layers are restricted by the rules of grammar.

Case based learning: when a good model cannot be established, through the record instance learning is a desirable method [6]. At this time, before the information is used, it does not make any processing. Using case based learning method, first of all, can be applied to any time compatible elicitation (consistency heuristic) method, the characteristics of a previously observed things assigned to another new things have never seen.

$$u'(x_1, x_2) = \sum_{s=-n}^n \sum_{t=-n}^n w(s, t) u(x_1 + s, x_2 + t) \quad (3)$$

Secondly, learn how to use the K- dimension tree quickly find the nearest neighbor object feature space. (PAST S1 KISS-ACTION AGENT (NAME J1 PERSON ("Jack"))) (THEM ENAME (NAME J2 PERSON "Jill")) it expresses a past case S1. PAST is an operator, said structure type is the past, S1 is the instance name, KISS-ACTION is the case form, AGENT and THEME is a description of the object, the agent and theme.

Assume that s and T are two different state of DFA M, we call s and T are equivalent: if s can read from the state of a word alpha and stop in the final state, then the same, starting from the t can read the same words alpha and stop in the final state; on the contrary, if the status of the t you can read a word alpha and stop in the final state, then the same, starting from the s can read the same words alpha and stop in the final state. If DFA M two state s and T are not equivalent, called the two state is distinguishing.

A grammar of G is defined as four tuple (VT, VN, S, R), wherein, VT is the end symbol set, is a non empty finite set; terminator is the basic symbols language. VN is a nonterminal symbol (or grammar, or set of variables), is a non empty finite set; the non terminal is used to represent the grammatical category; VT VN = phi. S called the recognition of symbols or start symbol. It is a conterminal symbol, at least as left appeared in a rule, as is shown by equation (4).

$$\tilde{g}(n) = (-1)^{n-1} h(-n + 1) \quad (4)$$

Automatic extraction of index terms, filtering, retrieval of natural language text, automatic extraction of important information, automatic abstract etc.. At the same time, because of the emphasis on "mass", stressed the "real text", the two fundamental work below has been paid attention to and strengthen the. Large scale through different depth of processing real text corpus, the research is based on natural language statistical properties. Without them, the statistical method can only be passive water.

The theory of natural language understanding, maybe is the keyword matching method, it is effective in some specific occasions [7]. The method is simple to sum up is this: the specified matching and action of two types of samples in the program? Then build a mapping action by matching to sample to sample. When the input statement with a matching to sample match, to perform the corresponding provisions of movements of samples, this seemingly machine truly realized the purpose of user questions can understand.

A simple sentence understanding does not involve the relation between sentences and sentences and understand it first Fu words with meaning, and then give the entire statement assigned to a structure. And a set of statements to understand, no matter it is an article excerpts or a dialogue excerpt, are required to discover the relationship between sentences.

IV. ANALYSIS OF NATURAL LANGUAGE UNDERSTANDING TECHNOLOGY BASED ON SEMANTIC WEB ONTOLOGY

There are many approaches for natural language understanding is a language generation, this is not only because they use different internal structure of expression (such as the use of semantic web or conceptual dependency etc.), and because the language generation different purposes (such as some purpose is to make the abstract, the input articles and some for as question answering machine interface system people - etc.). Language generation also has many difficulties, especially in the first part, more difficult. Sometimes, to exchange information by the question answering system can be estimated in answer to a question [8]. In these systems, the generation processes of information subject to answer the questions of the constraint.

Ontology is a formal, explicit specification of conceptualization of the domain knowledge, is to understand and describe the common domain knowledge. Ontology is the key to realize the semantic web technology; the ontology can solve the semantic description of the semantic web and the ambiguity of the two issues. The semantic web needs to achieve the number of ontologies greatly enriched and wide application, but now the WWW ontology is not only a very limited number.

The classification procedure thought. Tectonic type's defined task called concept learning. Of course, it used the technology dependent types (concept) description method. If is described by a score function, concept learning available parameter correction techniques [9]. If you want to structure type definition type, as is shown by equation (5).

$$\Sigma_N = \sigma^2 I_{(M-N) \times (M-N)} \quad (5)$$

Let G be a type 0 grammars, if produce each type of 61664 beta G all meet the $|\alpha\beta| \leq \epsilon$, except only s 61664 epsilon, but S may not appear in the right part of the production, then the grammar G is type 1 grammar or context sensitive grammar (CSG).

The definition of an equivalent: Let G be a type 0 grammar, if produce each type G are for $\alpha A \beta \Rightarrow \alpha \gamma \beta$, A in VN, and not the epsilon gamma,

alpha, beta, gamma epsilon (VT & VN) *, then the grammar G is type 1 grammar or context sensitive grammar.

The semantic web will enable human beings to liberation from search related Webpage heavy labor in. Because the network computer can use intelligent software, in search of tens of thousands of Webpage, and it is screening out the relevant and useful information through "intelligent agent".

Keyword matching method, although simple, but they ignore a lot of information in a sentence, to ensure that the sentence meaning details were not ignored, will determine which of the sentence structure in detail, it is to carry on the grammar analysis. To this end, we must first shows this specific language symbol string structure of the grammar, in order to comply with the rules of grammar sentence to produce a called the syntax analysis tree structure for each.

The database system is the effective organization and access most rely on the powerful database system of knowledge, can handle a large number of semantic object semantic web storage and retrieval problem, denoted by RDF or other markup language [10]. Scope of object semantic objects from the simple facts such as "John Smith: systems-analyst" to the complex such as e-commerce multi-agent negotiation protocol, as is shown by equation (6).

$$f_{resp}(x, y) = Det(Z) - kTrace^2(Z) \quad (6)$$

In the field of artificial intelligence (AI), the development and use of body language is a long-standing tradition, semantic web research can be the basis of further promoting the. At present, body language is the most important network in the field is as follows: XML provides a surface syntax for structured documents, but none of the meaning of a document to provide semantic constraints. XML Schema is a XML document structure definition language. A simple method of semantic analysis is to use semantic grammar. The so-called semantic grammar is based on the traditional phrase structure grammar, N (noun), V (verb) the concept of grammatical categories, with the discussion of specialized classes to replace the field.

RDF is very suitable to describe metadata information expression of Web resources, such as the title, author, date modified, copyright information, has the characteristics of simple, open, easy expansion, easy to exchange and easy integrated etc.. Because they are known as the Web resource, so RDF can actually describe any information can be identified on the network. Therefore in the resource description, RDF is more like a data model. The model describes network information resources in order to "resources - properties - property value" in the form of it.

XML reform information grammar representation way, to facilitate transplantation in different platform, language and application. XML has the ability of collaborative work of grammar, while DAML tries to collaborative work ability with semantic. The development of DAML due to a DARPA sponsored research projects; expand the XML so that it has a semantic relation and ability to express, "understand" that must be the semantic ontology knowledge. The key lies in the description of semantic

information in a way, in this way the aforementioned software agents can be seamlessly integrated into the Webpage, embedded Webpage software system, software system, and the interaction of the database.

In the semantic web to build a higher layer of a layer, we should follow two principles, downward compatibility. Located in a layer of the agent should be able to interpret and use the information in the low layer. For example, master OWL semantic agent must have sufficient capacity to make full use of, described by RDF and RDF Schema information. On the other hand, is located in a layer of the agent should be able to more high-level information using partially. For example, a master RDF and RDF only Schema semantic agent can be partly explained by the expression of OWL knowledge, while ignoring the other part on RDF and RDF Schema, the experimental results as shown in Fig .2.

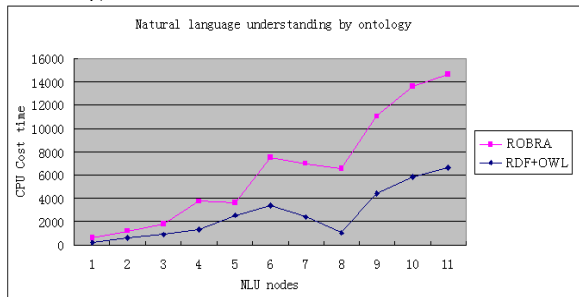


Figure 2. Comparison analysis of natural language understanding technology based on Semantic Web ontology.

To understand the system language part is composed of variable, format, a dictionary and grammar from natural language in Fig .2 of this proposed, linguists, they can be used to prepare the lexical analyzer. ROBRA is a tree diagram conversion system, used to complete the syntax analysis, transformation and generation. Language ROBRA is mainly composed of a control chart and conversion rules. The conversion rules define the subtree transform a variety of establishing multi-layer structure, through continuous transformation. One main purpose of the Semantic Web ontology is to specify the meaning of each item in a hidden code, so as to achieve the purpose of computer "understanding".

V. SUMMARY

Semantic analysis is through the analysis to find out the meaning, structure meaning and significance of the combination, so as to determine the true meaning or concept of the expression language. In automatic language understanding, semantic are increasingly becoming an important research content. This paper presents a natural language understanding technology of Semantic Web Ontology Based analysis. The core idea of the semantic web is to use a machine process able information network. Where the use of Ontology Web Language (such as OWL) is to describe the relationship between the clear implication and, it is to terms in the web documents and for them.

REFERENCES

- [1] Xin-hua Zhu, Qing-hua Cao, Fang-fang Su, "A Chinese Intelligent Question Answering System Based on Domain Ontology and Sentence Templates", JDCTA, Vol. 5, No. 11, pp. 158 ~ 165, 2011
- [2] Zhijuan Deng, Shaojun Zhong, "A Kind of Text Classification Design on the Basis of Natural Language Processing", IJACT, Vol. 5, No. 1, pp. 668 ~ 677, 2013.
- [3] LI Ning, XU Shoukun , LI Bo, Shi Lin, "An Efficient Ontology-based Semantic Web Services Composition Model for Peer to Peer Work", AISS, Vol. 4, No. 1, pp. 154 ~ 161, 2012.
- [4] Yi Wang, JieHong Luo, "Ontology learning and mapping in semantic web based on formal concept analysis technology", JCIT, Vol. 7, No. 10, pp. 381 ~ 388, 2012.
- [5] Lilac A. E. Al-Safadi , "Natural Language Processing for Conceptual Modeling ", JDCTA, Vol. 3, No. 3, pp. 47 ~ 59, 2009.
- [6] Chunlian Liu, Li Li, Yonglong Wang, "EA and Ontology to Use — A Study of Disaster Emergency Information Resource Catalog System (DEIRCS) for Resources Integration", JCIT, Vol. 8, No. 4, pp. 627 ~ 635, 2013.
- [7] Nurfadhilina Mohd Sharef, Shahrul Azman Noah, "Natural Language Query Translation for Semantic Search", JDCTA, Vol. 7, No. 13, pp. 53 ~ 63, 2013.
- [8] Xin-hua Zhu, Fang-fang Su, Qi-feng Tang, "A New Ontology-Based Semantic Similarity Algorithm in the Natural Language Processing", JDCTA, Vol. 6, No. 2, pp. 188 ~ 195, 2012.
- [9] Jinhyung, Myunggwon Hwang, Hanmin Jung, Won-Kyung Sung, "iLaw: Semantic Web Technology based Intelligent Legislation Supporting System", IJIPM, Vol. 3, No. 1, pp. 45 ~ 49, 2012
- [10] Pei Yin, Hongwei Wang, Wei Wang, "Extracting Features for Sentiment Classification: in the Perspective of Statistical Natural Language Processing", AISS, Vol. 4, No. 15, pp. 33 ~ 41, 2012