Research and Application of Web Information Retrieval Based on Ontology

Qi Shen, Meng Zhang, Qingming Song and Yan Tang

School of Software Engineering Beijing University of Technology, Beijing, China
shenq@bjtu.edu.cn, zhqmflag-2005@163.com

Abstract - This paper is mainly about the design of web information retrieval module based on ontology, using ontology technology, from the current web service mechanism based on the syntax level raised to the level of knowledge (or concept).First, the paper introduces the concept of the ontology, and then completes the architecture of the web information retrieval module based on ontology, with the technology of information extraction and semantic extension. This technology applies to web information retrieval, providing more accurate, intelligent web information retrieval service for the users, and improve the recall and precision ratio.

Index Terms - ontology; ontology reasoning; information extraction; semantic expansion.

1. Introduction

With the rapid development of the internet, web's development from web1.0 to web3.0.[1]Web2.0 is more focus on user's interaction, which are the visitor of the site's content, but also manufacturers of the site's content. Therefore, the information resources on the web are rapid growing and almost become a global information repository. Making the user become very difficult to query the information they want in this vast resource, A simple keyword matching mechanism is very difficult to be successful on improve the recall and precision ratio.

Web users usually organize information according to their own habits and knowledge, complex data structures is a major feature of the Web network, the data on the web is basically in the form of semi-structured HTML, it is lack of a description of the data itself, the semantic information is not clear, the pattern is also not clear, it is very difficult for the data query. Thus Web information extraction technology is appearing.

Semantic Web is a vision of the future network, the information is given explicit meaning, and the computer can automatically process and integrate information available on the Internet. Need is an ontology web Language (OWL) to describe the clear meaning of the term network documentation and the relationship between them. The ontology is to describe the conceptual model of the relationship between concepts. Performance concept hierarchy and semantic model as an effective, Ontology is widely applied to many areas of computer science. Has born many kinds of ontology languages[2], since the 90s of the last century, a number of ontology languages based on artificial intelligence have been proposed, such as KIF, Ontolingua, CycL, Loom, OCML and FLogic. With the development of Web-description language based on Web standards, such as SHOE (Simple HTML Ontology Extension), the XOL (XML based Ontology exchange Language), RDF, RDFS, OIL, DAML, DAML + OIL and OWL (Web Ontology Language)[.

2. Overview of the web information retrieval technology based on Ontology

A. The Web Ontology Description Language [3]

OWL is a web ontology representation language recommended by the W3C, it provides more formal semantic vocabulary. In the aspect of web content, computer's understanding is better than XML, RDF (Resource Description Framework) and RDF Schema (RDFS). The OWL contains 3 sub languages: OWLLite, OWLDL and OWLFull. OWL semantic elements includes class, individual, object attribute, data attribute, attribute characteristic, attribute constraints, Ontology mapping and the complexity class [4].

B. Semantic Representation of Ontology

Modeling method of ontology in a domain divide into five basic modeling primitives, that is a complete ontology should have 5 parts: class or concept, relation, function, axioms, Instances .From the semantic point of view, there are 4 basic relations: part-of relationship between the concept and expression of the kind-of expression; inheritance relationships between concepts, similar to object oriented in the parent class and subclass; instance-of relationship between the expression of concept instances and concepts, relations between objects and classes similar to object oriented; attribute-of expression to a concept is another concept attribute.

C. Methods of Ontology Construction

There are a lot of ways to build ontology, for example: the skeleton method, cyclic acquisition method, seven steps method, enterprise modeling method, IDEF5 method etc. And this article will use the 'seven steps method' to construct ontology. Using the ontology construction tool Protégé to build, construction ontology generally requires the following steps:

1) Determine which field it is belongs to: Determining the ontology belongs to a field, such as: tourist field, rather than geographic field.

2) Ontology reuse: Consider the use of the existing ontology.

3) Enumeration terminology: Determine the concepts and terminology of this field.

4) Define classes: To determine the relationship between classes.

5) Define attributes: The relationships between classes are connected by attributes.

6) Define attribute values: Attribute value is data attributes. Such as: bus starting station's name.

7) Define instance: Instance is the data of we really want to retrieve. Such as: attractions information, lodging information.

D. Semantic Expansion Algorithm

Semantic expansion algorithm, which is the concept extraction, uses keywords inputted by the user. It is according to the users semantic add the concept similarity retrieval words, and get more comprehensive, longer, more accurate retrieval set. The basic idea of the algorithm is: first, using the ontology to make the keywords standardization inputted by user, transformed into a standard concept description. The second is an extension of the concept, using the ontology hyponymy. The extended concept algorithm, using the concept of hierarchy, will widen the formation of concept, concept retrieval set. Finally, this module compares the set with the ontology library, query the concept was not there originally. Semantic expansion algorithm is described as follows [6]:

Users enter keywords: K

Output expansion concept collection: Q

User input keywords with ontology library standardization, initialize Q = null.

K and K synonymous concept add to the set Q of concept, to generate a new set of concepts.

If K is a class concept, jump to d); Otherwise, K belongs to the class and instance concept of K added to Q as Search conditions. If the result is empty, goto d); otherwise returns the query results, query end.

K direct parent class and indirect parent class added to the Q set, K sub-class instance added to Q, then continue as the search condition query, if the result is empty, goto e); otherwise return a result set to the end of the query.

Each associated node and then iterate through the breadth-first algorithm or a depth-first algorithm, the initial set of the next query is generated.

So that after semantic expanded to produce a new set of concepts, no longer a simple one or a few keywords, to finish the ultimate collection of concept in the result set.

E. Information Extraction

Information extraction (IE) is processing the information contained in the text into structured, and then turn into a form of the same organizational form. Web information extraction method has many kinds of different methods [7]:

1) Machine learning information extraction based on the technology of information extraction: This way is based on the delimiter to locate to extract data. According to the advance by the user labeled examples to automatically learn and generate extraction rules based on the delimiter.

2) The information extraction of natural language understanding based on the technology of information extraction: The way of the commonly used for information extraction from free text, use the words, phrases and establish the relationship between the structure of sentence extraction rules based on syntax and semantics, in order to achieve information extraction.

3) The information extraction based on Ontology type: Information extraction technology the way mainly uses the description information of the data itself to achieve data extraction, mainly rely on a complete knowledge base. The knowledge base is defined between the extraction pattern of each element and their relationship, this kind of information extraction is not dependent on the webpage structure and form.

4) The information extraction based on HTML structure: Information extraction technology is based on the structure of the HTML positioning information, before information extraction using a parser to parsing the Web document into a tree model, extraction rules generated by automatic or semi-automatic way; information extraction is converted to extract information from the syntax tree of the operation.

3. Design and Implementation of the web information retrieval module

A. Determine the Ontology Belongs to Which Fields

Different areas to create the ontology, included in the concept of instances, properties may be different. For example: a scenic Name the Summer Palace, according the field of tourism to build ontology, the ontology may want to include information about attractions around route, surrounding accommodation costs and other related information. To build ontology by geographical area, the ontology may not mention above. Build ontology library, we must first determine the ontology belongs to which fields.

B. Ontology Encoding

Domain ontology concepts are determined, next is ontology encoding, ontology encoding stage include: the choice of ontology language, the choice of ontology development tools. And then the bulk of the coding works that ontology concept to describe, to establish the relationship between related instances, attributes and classes.

TABLE I Ontology Development Tools Comparison Table

Name / Function	Protégé	Webonto	OntoEdit	WebODE	OILED	Ontosaurus	Ontolingua
OWL	√			√	√		
Visualization	√	√		√			
Ontology merging	√		√	√			√
Chinese support	√						
Network technology		√	√	√		√	√
Fuzzy Ontology							
Cooperative development		√	√	√		√	√

From table 1 finally choice Protégé ontology development tool[5]. Protégé ontology editor tool developed by Stanford University's medical information study group, it is a free and open-source platform, and it can use RDF, RDFS, OWL ontology description language editing and modify the body.

Provide a reliable and efficient environment for the development of the ontology.

C. Ontology Reasoning and Querying

For ontology reasoning work mainly through the Jena development package to achieve, where Jena is the rules-based inference engine, the establishment of rules is pushed well into the Jena reasoning, reasoning based on ontology, all the implicit description information can be display, more and more information is returned to the users [8, 9].

Inference rule based on ontology is composed by many rules. Each rule is composed of a main body and a head, a rule can have a body and a head, for example: Rule1: (? X hasStation? Y), (HasBus? Y? Z), (isAccessible? Z? W) → (isAccessible? X? W), the main body of rules: (hasStation? X? Y), (HasBus? Y? Z), (isAccessible? Z? W). the head is: (isAccessible? X? W), meaning that all entities can reason the head (? X hasStation? Y), (HasBus? Y? Z), (isAccessible? Z? W), (isAccessible? X? W). they have one name: ClauseEntry. For example, methods in the Rule class getbody () method returns a ClauseEntry set. He has 3 elements (hasStation? X? Y), (HasBus? Y? Z), (isAccessible? Z? W). To establish the reasoning rules, the relevant knowledge can learn Jena inference machine.

Ontology and inference rules based on ontology components the entire data set. Contains semantic information, through this collection can be achieved on the user selected need to show the semantic level, the above process, namely ontology query. Therefore, using the query language (for example SPARQL query language) in the ontology condition, make search on the Internet to preprocess the data, and then the retrieval results are stored in the database, so as to realize the information extraction semantic hierarchy. Simple said information extraction system is a process of establishing the database resources, also can be the process of query Web information first, and then put the organized structure of data stored in the database.

D. Design of Information Retrieval Module

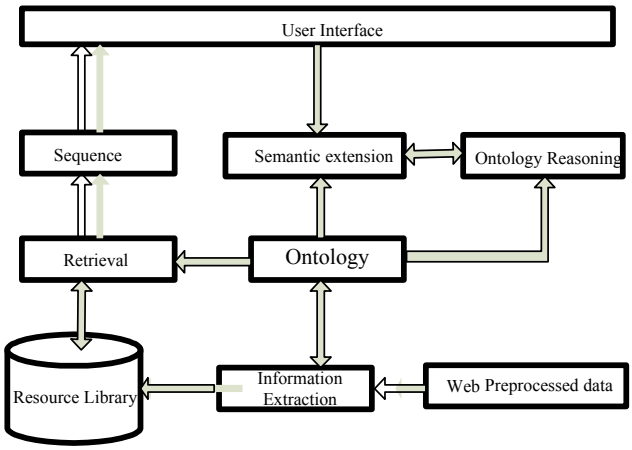

Figure 1 Ontology-Based Information Retrieval Chart

1) User interface: the user through the user interface submits a query, the user interface‘s mainly function is interact with the user. When the user is presented with a retrieval request, submit it to the query analyzer. When retrieving information is completed, the retrieval results returned to the user.

2) Semantic extension: analysis the user's request, including the analysis of semantic and semantic extension, which should be based on ontology.

3) Ontology reasoning: Using Jena framework combined with domain ontology reasoning, expanding the query conditions.

4) Retrieval: Query resource library to find all related documents, finally sorting module according to the correlation between the document and user’s query requests, and then sort the documents, Returned to the user in order.

5) Ontology: Ontology library is used to store the ontology model. At the same time, because the ontology has to share, to reuse, analysis of knowledge, knowledge acquisition, knowledge of the role of standardization, so query analyzer, the reasoned, retrieval and information extraction also needs ontology to complete.

6) Information extraction: under the action of ontology, make the pretreatment of web resources into structured, machine-understandable resource library.

4. Conclusion

This paper introduces the ontology and ontology expansion algorithm, which uses the ontology technology applying to web information retrieval. Jena which is a development framework, implements web information retrieval based on ontology. The semantic retrieval solves the problem of heterogeneous data, and improves the recall ratio and the precision of web information retrieval.

Although the research of web information extraction technology makes some progress, it is still a new field to be studied in the stage of exploration. The first problem to be solved is to construct high quality ontology. The ontology design is a creative process. However, domain ontology construction is a very challenging task. Firstly, the construction of ontology requires the experts in the field. The second job is the pretreatment of web information. So it needs further efforts to improve the construction of ontology, complete ontology rules. The search mechanism in semantic level will be better.

Further research of the building of domain ontology and ontology reasoning will be done. It combines with the feasibility and rationality of information services and e-commerce to apply in scientific research and daily life, so it brings convenience for research workers and people’s life.

5. Acknowledgment

This work is supported by Scientific Research Project of Beijing Municipal Commission of Education (Grant No. KM201210005030), the support is gratefully acknowledged.

6. References

[1] Ying Zhang Web3.0: personalized learning platform [J]. China education technology equipment. 2012 (27): 38-39.

[2] Zhihong Deng, Shiwei Tang, et al. Overview of Ontology [J]. Journal of Peking University: Natural Science Edition, 2002, 38 (5): 730-738.

[3] Jianhou Gan , Youming Xia, Tianren Xu, et al. Ontology knowledge representation extends the [J]. language OWL Journal of Yunnan Normal University (NATURAL SCIENCE EDITION). 2005 (04).

[4] Binbin YU and construction tools of [J]. border economic and cultural ontology construction method. 2012 (12): 167-168.

[5] Yong Zhang, Junbai Lv Protege Ontology Modeling Research Based on [J]. Fujian computer. 2011 (01).

[6] Tao, Teng-Yang, Zhao. An Ontology-Based Information Retrieval Model for Vegetables E-Commerce[J]. 2012, 11 (5): 800-807.

[7] Chengyi Che, Zongmin Ma, Xiaolong Jiao . Study on the recognition method of [J]. computer engineering data table in the Web page. 2012, 38 (23): 154-157.

[8] Hong Tian, Pengyun Ma. Jena city transportation domain ontology inference and query method based on [J]. computer applications and software. 2011, 28 (8): 57-59.

[9] He, Youquan, Xu. Design and Implementation of Ontology-Based Topic Retrieval System[J]. 2012, 9 (5): 523-529