ICIC Express Letters ICIC International ©2016 ISSN 1881-803X
Volume **10**, Number **9**, September **2016** pp. 2171–2175

EXPERIMENTAL STUDY ON THE CONSTRUCTION OF DERMATOLOGY'S SEMANTIC METADATA

Jianping Zhang[1] and Yao Liu[2]

[1]Medical Library of the Chinese PLA
No. 59, West Fourth Ring Road, Beijing 100039, P. R. China
zjping51@126.com

[2]Institute of Scientific and Technical Information of China
No. 15, Fuxing Street, Beijing 100038, P. R. China
liuy@istic.ac.cn

Received February 2016; accepted May 2016

Abstract. *Based on the theory of "interactive of content and style of library resources organization's semantic methods", this study is conducted with traditional resource organization methods and label ontology in the platform of ontology construction and MDC. Skin disease is taken as an example to provide theoretical and practical bases for the semantic research, so as to construct a dynamic development system of ontology for dermatology. This test also stores the outcomes of semantic annotation in the form of semantic indexing, using the indexes through a retrial of asking and answering questions.*

Keywords: Ontology, Metadata, Semantic annotation, Skin diseases, Natural language processing

1. Introduction.

Semantic metadata is a conceptual model that can describe the information system at the semantic and knowledge levels. As it supports information exchange and knowledge sharing between human and machine, it has been widely reused as an ideal data model [1]. Based on the creation of ontology knowledge, semantic metadata involves not only the use of ontology entities, but also the collection of annotated domain knowledge [2]. In this period, both ontology research theory and method need to be further improved. The reason lies in that many ontology entities are developed artificially at the present stage, which will not only cost a lot of manpower, material and financial resources, but also consume a very long time [3]. At present, the automatic ontology constructions are rare for natural science, and inexistent for skin disease, besides some small experiments reported in studies with natural language processing technology [4]. Besides, the semantic metadata in the field of dermatology is also blank. In this experiment, a Multi-Domain Corpus Annotation (MDCA) system is utilized to generate a domain dictionary, and complete the functions of dictionary editing and modification. The obtained domain dictionary is then used to cut and annotate the raw texts [5]. With traditional resource organization methods and label ontology, an ontology for dermatology is constructed on the ontology construction platform, which is then left to evolve automatically and published in the form of Wikipedia to facilitate the professors to edit it online simultaneously. The results of the semantic annotation are stored in the form of indexes and reused. The indexes of semantic annotation are used as domain knowledge for retrieval of question answering system [6].

2. Knowledge Preparation.

2.1. Semantic dictionary of dermatology.

The word "semantic" means that the texts should be understood by a computer. This is closely related to the natural language processing technology, which depends on the size and quality of the corpus. As only a list of words with a certain rule, an ordinary dictionary cannot reflect the relationship between the entries. However, the professional semantic dictionary not only contains a list of professional terms, but also presents the logical relationship among them [7].

The generation tool for semantic dictionary is supported by the MDCA system, which is a professional semantic dictionary generation tool flat2tree.jar. Users only need to upload the required file format in accordance with the system requirements, and the system can automatically generate a professional semantic dictionary. The input texts used to generate the semantic dictionary shall meet the requirements of the input system. The format is as follows. Each line has two terms with a mutual relationship like that between a father and a son. The former one is the father, and the latter one is the son. The two words are separated by a space. The generated semantic dictionary format is presented below. Two terms with a mutual relationship as that between a father and a son are divided by "Tab" key. Each term is annotated by the acronyms. As the dictionary is generated by the system, the system will activate the functions associated with the professional semantic dictionary, and the generation function of the dictionary will not be used. At this point, the function of the professional semantic dictionary is to view the semantic tree, save the database, and export professional semantic dictionary.

MDCA system supports two kinds of editing forms of the dictionary, namely attribute and content editing. Attribute editing is the editing of the classification and authority information, which means that the dictionary authority is either open or private. A public dictionary can be viewed by users, and applied as an annotated tool. Users can also edit and modify the open dictionary, and then store it as a private dictionary. When the dictionary is saved as a private dictionary, the system will check users' dictionary classification information. If the users have not exited from the dictionary, the system will save it as an updating of their dictionaries. Otherwise, the system will not modify the dictionaries. Content editing is a modification of the dictionary tree. In users' browser, a dictionary is in the form of a tree, and a node on the tree is a term of the dictionary. After clicking on the term, a modification option will emerge, such as "add a sub-level term", "add a same-level term", "delete the term", and "modify the term". According to the system settings, users can modify the term in these four forms. At the same time, the root node of each dictionary tree is named "JSON", and all the terms in the dictionary are the sub-term of this node. Users cannot delete and modify the root node, but they can add a sub-level term on condition that some requirements for the format are met. The format should satisfy the system's requirements and term annotation. If the users add the same term at the same level, the system will automatically detect it and give a hint to the users, and then add a number to the annotated term to show the difference.

2.2. **Semantic annotation of dermatology knowledge.** Semantic annotation of documents is the basic work of natural language processing. This study is to use WEB (MDCA) developed by this research group. The overall framework of the technology is based on the mutual promotion between dictionary generation and semantic annotation. When users upload or input texts, the MDCA system would activate word cutting component, and load the professional semantic dictionary at the same time. MDCA system is developed after a full investigation and experiment of Chinese segmentation and tagging technology. On the basis of Python technology, the Jieba is the core of the Chinese open source segmentation technology. The workflow is based on the PHP technology. The entry of the annotated reference is Liu Yao's "Chinese medicine ancient literature corpus design and development research" referring to the semantic type of naming rules. The system's cutting and annotation process of the text is divided into three steps: inputting, processing, and outputting. When uploading texts to the server, the users select a proper professional semantic dictionary, and input or upload word cutting and annotation component, and then the system would cut and annotate the texts.

For the annotation results, the MDCA system also has a unique processing. Users upload TXT format file, and the system generates a tree structure of professional semantic dictionary. The system shows performances for professional semantic dictionary file (.Txt), and professional semantic dictionary file JSON format file (.Json), based on the hierdict pickle file (based on the Pickle Python data format dictionary cache file), and the result file (.Posed).

3. Construction of Semantic Knowledge for Dermatology.

3.1. **Generation of semantic metadata for dermatology.** The key to constructing the semantic system is to access the concepts and the relationship among the concepts, and transform them to the form that the ontology can describe. In this construction work, MESH is chosen to describe the ontology relationship, by referring to the professional textbook and dictionary on dermatology to set the properties of concepts. Mesh terms' classification and property settings should be adjusted and modified in order to build the semantic system with each term containing only one ontology element. On the ontology construction platform, the prepared materials are input into the system to construct the ontology for dermatology automatically. Some other functions are also accomplished automatically, such as import of properties, hierarchy generation, extraction of ontology terms' relationship, and ontology evolution.

3.2. **Automatic evolution of semantic metadata.** Ontology is a description of the relationship among concepts. As everything changes, ontology also needs to have adaptability. For the evolution of ontology, Internet data are chosen for the characteristic that the point of view is relatively new. Nevertheless, Internet data have network noise. This disadvantage is also its advantage. If one uses the same key words to query, the results are mostly related to some conclusions, which most people approve. It is possible that these conclusions are likely to be correct. Besides, some of the results are discarded because of network noise, which will not cause a great impact on the overall result. Therefore, Internet data are selected for the evolution of ontology.

Ontology construction platform supports two kinds of automatic evolution methods. One is systematic evolution, and the other is personality evolution. The operation for systematic automatic evolution is very simple. Users only need to choose the class that needs to evolve, and then click on the automatic evolution button on the toolbar of the platform. Subsequently, the system presents a prompt. If users want to choose the type of evolution, they just have to click "Yes". For personality evolution, users first need to specify the evolution conditions, and then set the value of the professional attributes of the class. As a complement for systematic evolution, personality evolution supports four types: the specified classes, attributes, keywords and sites. For example, the seed file format to contact dermatitis is shown below: contact dermatitis: diagnosis = patch test; diagnosis; etiology, pathogenesis and pathological = plant; epidermis; 2 = hormone therapy, calamine; clinical manifestations of 2 shares, eczema.

The learning site is "Good Doctor Website" which has information about dermatitis. For selecting personality evolution of the platform, the system will pop up options for users to choose, such as the big class and seed file. After users input learning site and click evolution type, the system will start, and evolution results would be presented in the record.

3.3. **Study of semantic metadata.** A self-learning mechanism is developed for the frequently updated features of networking domain semantic resources. According to the characteristics of the domain resource, the URL seed collection areas of the site, as well as the crawl's timing, frequency and other parameters are set. A domain network crawler

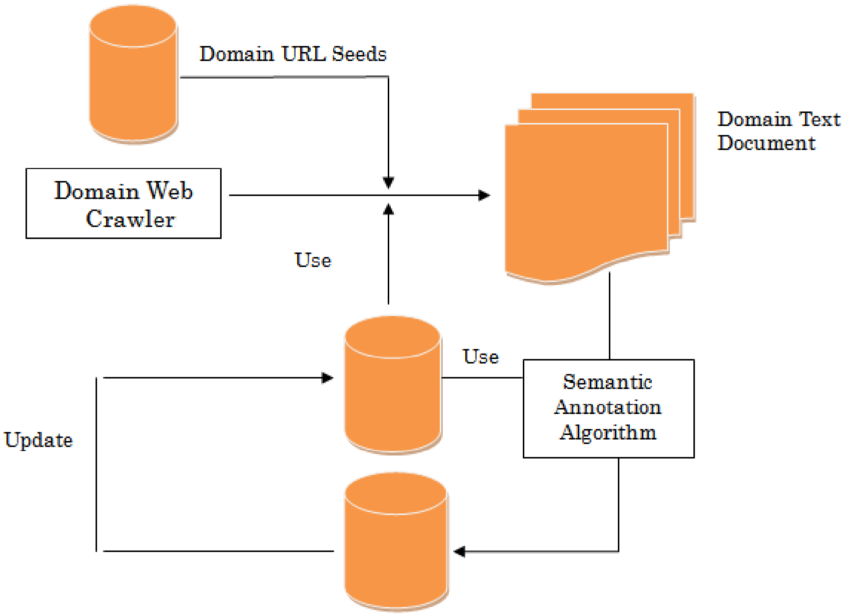

Figure 1. Resource learning algorithm

is applied to crawl text field knowledge to annotate semantic data, and effective domain data are adopted to update the domain semantic indexing fields.

The approach of crawling the updated network resource is adopted to update exiting semantic repository, while high-quality resources are utilized to update the semantic ontology repository. Then, the updated ontology repository is made use of to crawl the updated resource from the network, or annotate semantic resources and so on to form a storage platform for semantic resources to build a semantic resource learning mechanism automatically.

If users want to use semantic metadata for study, they should set the range of definitional domain at first on the auxiliary construction platform. Then, they shall click the function keys, select the name of the project, the class to learn, and data files, and then learn the materials by attributes in the attribute column. The learnt results of the properties appear in Properties 2. Subsequently, learning function is to be accomplished. Based on the above process, the platform can study other learning contents from other documents. The document appears in the properties, and the learnt results of the properties appear in Properties 2.

4. Application of Semantic Metadata.

4.1. Publication of semantic metadata.

After the completion of ontology construction, the obtained ontology still needs to be modified. In this modification process, domain experts play an important role. In order to facilitate the domain experts to modify the ontology online and support a number of domain experts to edit online, the platform will publish the ontology in the form of Wikipedia. Wikipedia is a free content for the purpose of collaborative editing, because of its simple interface and popularity among people all over the world. As a result, the platform chooses to publish the ontology in the form of Wikipedia online. This function has been integrated into the platform system.

The platform supports the publication of domain ontologies in the form of Wikipedia, and the publication category can also be chosen in the domain ontology. The default of the system is that the sub-classes will be published along with the parent class. After the domain ontology is published, the domain ontology is synchronized with the entries in the Wikipedia. All the changes in the domain ontology are synchronized with the Wikipedia entry.

4.2. **Retrial system for asking and answering questions.** The query system based on the data of the construction platform is constructed with the aid of semantic metadata. The function of the system users in the query interface is to key into search terms, and then the system will compare the data with the back-end storage after the transfer of returns in the ontology classes and the attributes of the content. The class is returned to display the position relationship in the domain ontology, and the content of the attribute is that of the semantic index after the semantic annotation. Taking skin disease for example, users input contact dermatitis, and contact dermatitis and its properties in the ontology database and attribute values will be presented.

Semantic metadata knowledge retrieval and question answering is a semantic indexing content based on ontology data and semantic annotation of domain ontology. Users query the search terms in the search interface, so that the system will extract the background data, and then return the relevant content.

5. **Conclusions.** In this study, ontology construction platform and MDCA system are adopted to carry out a series of experimental studies with traditional information management methods, aiming to construct a semantic metadata for dermatology. The domain dictionary is generated and used to annotate the texts. Indexes of semantic annotation are applied in the knowledge question answering system to retrieve the field of skin diseases, which has achieved a good result. In consequence, this ontology construction method is relatively suitable for the semantics of domain knowledge. Because of the searching engine, the semantic annotation is not sufficient. In the future, more learning materials will be used in the construction system to update the system.

Acknowledgment. This study is partially supported by Zhijun Guo and Yi Huang. The authors are also grateful for the helpful comments and suggestions from the reviewers, which have improved the presentation.

REFERENCES

[1] Y. Liu, Z. Sui, Y. Hu et al., Research based on the theory and method of library resources organization semantization, *Information Studies: Theory & Application*, vol.33, no.10, pp.105-106, 2010.

[2] Y. Liu, Z. Sui, Y. Hu et al., Automatic construction of domain ontology, *Journal of Beijing University of Posts and Telecommunications*, vol.29, no.Z2, pp.65-69, 2006.

[3] Y. Liu, Research of approaches and development tools in constructing ontology, *Journal of Modern Information*, vol.29, no.9, pp.17-20, 2009.

[4] Z. Guo, *Research and Develop of Ontology Construction Platform*, Peking University, 2013.

[5] Z. Sui, Y. Liu and Y. Hu, Extracting hyponymy relation between Chinese terms based on term types' commonality and sequential patterns, *ICIC Express Letters*, vol.3, no.4(B), pp.1233-1238, 2009.

[6] D. Zheng, *Research of Design and Complete A Platform Based on Web Semantic Data*, Peking University, 2013.

[7] Z. Xiao, *Research of Design and Complete A System Based on Web-based Multi-domain Corpus Annotation System*, Peking University, 2013.