Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Report:

National semantic infrastructure for traditional Chinese medicine

Hua-jun CHEN

School of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China

E-mail: huajunsir@zju.edu.cn

doi:10.1631/jzus.C1101012

1 Introduction

Traditional Chinese medicine (TCM) is "the traditional medicine that originated in China, and is characterized by holism and treatment based on pattern identification/syndrome differentiation" (WHO, 2007). In China, traditional medicine accounts for around 40% of all healthcare delivered; in the West, TCM is increasingly adopted by medical practitioners as a form of complementary and alternative medicine (CAM) (WHO, 2002). Both in China and in the West, scientists are attempting to bring the ancient heritage of TCM into line with modern standards, through the scientific development of TCM in the direction of evidence-based medicine (Qiu, 2007). There is a pressing need for the digital preservation and global access of TCM knowledge assets.

1.1 Digitization of TCM knowledge assets

The major approach to the digitization of TCM knowledge assets is to engineer TCM databases. For example, Fang *et al*. (2008) generated a database through text mining, which captures associations between TCM, gene, and disease information. Bensoussan *et al*. (2002) developed a database to collect scientific evidence on the toxicity of Chinese herbal medicine. Notably, China Academy of Chinese Medical Sciences (CACMS) has taken a series of efforts in TCM database construction, resulting in the Traditional Chinese Medicine Database System (TCMDBS), which "consists of over 40 categories of Chinese medicine databases, possessing 1 100 000 items of data" (http://cowork.cintcm.com/engine/windex1.jsp). These databases can be utilized using computational methods applied in TCM use cases such as genetics, studies of mechanism, prescription validation, and herbal medicine toxicology.

Currently, the major challenge facing TCM informatics is the 'data island' problem. The digitalized knowledge assets, though large in quantity, are distributed among various databases and systems which cannot be easily connected. The obstacles to knowledge integration limit the effectiveness of knowledge discovery in integrative studies.

1.2 Semantic Web in action

The Semantic Web, also known as the Web of data, is an emerging technical solution to the 'data island' problem. Berners-Lee *et al*. (2001) unveiled a vision of the Semantic Web: an interconnected Web of data that could be easily accessed and understood by humans as well as the machine. To realize this vision, the Semantic Web community has contributed a bunch of technologies, including the Resource Description Framework (RDF), expressive ontology, Web-accessible repositories, query and reasoning engines, and information extraction techniques. As Semantic Web technologies are reaching maturity, they are applied in various domains such as biology, medical sciences, and healthcare.

We have conducted the TCM-Grid project, supported by National Key Grid programs and National 973 (Basic Research Program) Semantic Grid programs. In this project, we utilized Semantic Web technologies to integrate the heterogeneous databases with the TCMDBS system and developed the world-largest semantic infrastructure that provides a rich set of knowledge services to Web users within the TCM community worldwide. It is a living

embodiment of the 'Knowledge as a Service (KaaS)' model, which emphasizes providing the TCM knowledge assets as open knowledge services. By implementing the KaaS model, TCM knowledge assets are no longer locked inside isolated databases or libraries, and become Web resources that can be easily accessed, combined, and reused.

2 System architecture

We use a domain ontology to construct a Semantic Web environment to unify and link the legacy databases, which typically have heterogeneous logic structures and physical properties. Users need only to interact with the Semantic Web environment, and perform searching, querying, and navigating around an extensible set of databases without the awareness of the database boundaries. Additional deductive capabilities can then be implemented to increase the usability and re-usability of data. According to Wu and Chen (2009), the system contains three layers: resource layer, service layer, and application layer.

The resource layer provides a uniform mechanism for the management of a plurality of resources in the TCM domain, and supports remote operations on the contents of resources.

The service layer provides the following services for knowledge manipulation and integration:

1. The process semantic service exports process semantics as Ontology Web Language for Services (OWL-S) descriptions.

2. The database semantic service publishes the schema of relational databases as semantic description.

3. The ontology service exposes the shared domain ontology and provides basic operations on the ontology.

4. The semantic mapping service facilitates users to edit the mappings between local resources and the mediated ontology.

5. The semantic query service answers semantic queries by rewriting them into Structured Query Language (SQL) queries, and wrapping the results of SQL queries into RDF triples.

6. The semantic search service provides semantic-based full-text search based on databases.

The application layer contains a rich set of Semantic Web applications, such as ontology engineering, knowledge acquisition, database integration, information retrieval, and knowledge discovery, which will be introduced in the next section.

We have given a more detailed description of the system architecture in Wu and Chen (2009).

3 Applications

In the DartGrid project, we focus on three major TCM requirements, including academic virtual organization, personalized healthcare, and drug discovery and safety (Wu and Chen, 2009). Here we present a brief overview of the major applications that we have developed to satisfy the above requirements.

3.1 DartOnto: ontology engineering for TCM

DartOnto is a Web-based tool for ontology engineering and knowledge acquisition. We have invented a generic methodology (model, process, and tools) that both fits the problems in our project and applies to other situations (mainly about public services sections, including healthcare, life sciences, knowledge assets, and sociology). The methodology focuses originally on the TCM pharmaceutics domain (the knowledge about herbs, herbal formulae, and their characteristics), and should be scalable upwards to the entire coverage of TCM knowledge, as well as downwards to the topic level. Specifically, the methodology is decoupled with specific ontologies and data used in our project, and DartOnto incorporates a set of generic semantic techniques optimized for the methodology.

With DartOnto, we have built a TCM domain ontology named the Unified Traditional Chinese Medicine Language System (UTCMLS), which is the largest TCM ontology and achieves a comprehensive coverage of the TCM domain (Feng *et al*., 2006). The UTCMLS ontology has been used for the acquisition of TCM knowledge assets, through manual editing of semantic annotations, and automatic methods for information extraction.

3.2 DartMapping: semantic integration platform

We have built a semantic data space based on the UTCMLS ontology. The data space achieves the semantic integration of relational databases in a pay-as-you-go manner, by mapping the data within

relational databases into virtual RDF graphs that are accessible via SPARQL queries. To achieve relational-database-to-RDF (D2R) mapping, we developed (1) a tool that defines the mapping rules between relational schemas and the shared domain ontology and (2) a query-rewriting engine that translates a SPARQL query into a series of SQL queries against underlying relational schemas according to the predefined mapping rules. As a result, we integrate the databases within the TCMDBS system, and provide a Web portal for unified and coherent access of TCM knowledge assets.

3.3 DartSearch: intelligent semantic search engine

This application enables the search of knowledge assets and full-text documents using keywords, concepts, and semantic links. It follows a similar interaction paradigm to existing search engines such as Google and Yahoo: a user enters keywords related to the interested concepts (item, category, or topic), and the system returns a ranked list of results that may be semantically relevant to the query from multiple sources. This application also enables Web users to navigate the knowledge space as a visualized graph of things.

3.4 DartQuery

This application facilitates complex question-answering through semantic queries in terms of the TCM ontology. The user interface enables users to compose SPARQL queries and view query results in tables or visualized graphs. It enables the construction of a SPARQL query in two modes, free mode and aided mode. In the latter mode, the system guides a user through the construction and refinement of a SPARQL query in an incremental process.

3.5 DartSpora: mining the Web of data

This application facilitates users to perform knowledge discovery experiments on the Web of data. We proposed a methodology named semantic graph mining (SGM), which integrates graph mining with ontology reasoning for mining on the Semantic Web. Intelligent agents, empowered with SGM, can predict hidden links, identify frequently occurring patterns, and learn interesting rules based on the Web of data. This service enables a variety of use cases that distill valuable insights from TCM data, two of which are explained as follows:

1. Analyzing the complex networks in the TCM domain. Complex networks capture the structure and dynamics of complex systems in the TCM domain. Here, we focus on the analysis of TCM complex networks, which brings new understanding of and insights into the complex systems in the TCM domain. For example, it is possible to correlate ethnomedical use with experimental biochemical or pharmacologic activities to identify plants having both types of activity for a given effect. Therefore, a network of herbs can be mapped to analyze herb-drug interactions and learn rules for formula combination.

2. Analyzing the collaborative network of TCM practitioners. This study uses the TCM literature to map a collaborative network of TCM practitioners by associating authors and topics within documents. The major relationship under investigation is collaborative authoring, and the data model is an expert-document-topic triple. This network is used to characterize the collaboration among TCM practitioners.

4 Conclusions

Our system can serve TCM practitioners, physicians, and scientists worldwide, with knowledge assets covering a broad range of topics including basic theories, diagnostics, diseases, therapeutics, acupuncture and moxibustion, and medicinal treatments. The TCM Search project demonstrates the maturity and advantages of Semantic Web technologies, and contributes to the preservation of TCM cultural heritage and the promotion of cross-cultural and interdisciplinary dialogues.

References

Bensoussan, A., Myers, S., Drew, A., Whyte, I., Dawson, A., 2002. Development of a Chinese herbal medicine toxicology database. *Clin. Toxicol.*, **40**(2):159-167. [doi:10.1081/CLT-120004404]

Berners-Lee, T., Hendler, J., Lassilia, O., 2001. The Semantic Web. *Sci. Am.*, **284**(5):34-44. [doi:10.1038/scientificamerican0501-34]

Fang, Y.C., Huang, H.C., Chen, H.H., Juan, H.F., 2008. TCMGene DIT: a database for associated traditional Chinese medicine, gene and disease information using

text mining. *BMC Complem. Altern. Med.*, **8**(1):58. [doi:10.1186/1472-6882-8-58]

Feng, Y., Wu, Z., Zhou, X., Fan, W., 2006. Knowledge discovery in traditional Chinese medicine: state of the art and perspectives. *Artif. Intell. Med.*, **38**(3):219-236. [doi:10.1016/j.artmed.2006.07.005]

Qiu, J., 2007. China plans to modernize traditional medicine. *Nature*, **446**(7136):590-591. [doi:10.1038/446590a]

World Health Organization (WHO), 2002. Traditional Medicine Strategy. Available from http://www.who.int/medicines/publications/traditionalpolicy/en/index.html

World Health Organization (WHO), 2007. International Standard Terminologies on Traditional Medicine in the Western Pacific Region. Available from http://apps.who.int/bookorders/anglais/detart1.jsp?codlan=1&codcol=52&codcch=2107

Wu, Z.H., Chen, H.J., 2009. Semantic Grid Applications for Traditional Chinese Medicine. *In*: Semantic Grid: Model, Methodology, and Applications. Zhejiang University Press, Hangzhou, China, and Springer-Verlag GmbH Berlin Heidelberg, p.195-209.

Recommended reading

Chen, H., Wu, Z., Mao, Y., Zheng, G., 2006. DartGrid: a semantic infrastructure for building database grid applications. *Concurr. Comput. Pract. Exper.*, **18**(14):1811-1828. [doi:10.1002/cpe.1031]

Chen, H., Mao, Y., Zheng, X., Cui, M., Feng, Y., Deng, S., Yin, A., Zhou, C., Tang, J., Jiang, X., *et al.*, 2007. Towards Semantic e-Science for traditional Chinese medicine. *BMC Bioinform.*, **8**(suppl 3):S6. [doi:10.1186/1471-2105-8-S3-S6]

Chen, H., Ding, L., Wu, Z., Yu, T., Dhanapalan, L., Chen, J.Y., 2009. Semantic Web for integrated network analysis in biomedicine. *Brief. Bioinform.*, **10**(2):177-192. [doi:10.1093/bib/bbp002]

Wu, Z., Chen, H., 2008. Semantic Grid: Model, Methodology, and Applications. Zhejiang University Press, Hangzhou, China, and Springer-Verlag GmbH Berlin Heidelberg.

Feng, Y., Wu, Z., Zhou, X., Fan, W., 2006. Knowledge discovery in traditional Chinese medicine: state of the art and perspectives. *Artif. Intell. Med.*, **38**(3):219-236. [doi:10.1016/j.artmed.2006.07.005]