Information Retrieval and Knowledge Discovery on the Semantic Web of Traditional Chinese Medicine

Zhaohui Wu
College of Computer Science,
Zhejiang University,
Hangzhou, 310027, China
wzh@zju.edu.cn

Tong Yu
College of Computer Science,
Zhejiang University,
Hangzhou, 310027, China
ytcs@zju.edu.cn

Huajun Chen
College of Computer Science,
Zhejiang University,
Hangzhou, 310027, China
huajunsir@zju.edu.cn

ABSTRACT

We conduct the first systematical adoption of the Semantic Web solution in the integration, management, and utilization of TCM information and knowledge resources. As the results, the largest TCM Semantic Web ontology is engineered as the uniform knowledge representation mechanism; the ontology-based query and search engine is deployed, mapping legacy and heterogeneous relational databases to the Semantic Web layer for query and search across database boundaries; the first global herb-drug interaction network is mapped through semantic integration, and the semantic graph mining methodology is implemented for discovering and interpreting interesting patterns from this network. The platform and underlying methodology are proved effective in TCM-related drug usage, discovery, and safety analysis.

Categories and Subject Descriptors: H.4.m [**Information Systems**]: Miscellaneous

General Terms: Algorithms, Experimentation, Human Factors.

Keywords: Semantic Web, Information Retrieval, Knowledge discovery.

1. INTRODUCTION

The health care and life sciences communities have already taken efforts in the adoption of Semantic Web technologies, including ontology engineering, semantic integration, information retrieval, and knowledge discovery. However, these successful projects focus exclusively on orthodoxy medicine, and never on Traditional Chinese Medicine (TCM) domain. Indeed, despite its wide adoption in Chinese communities, TCM has rarely been the application domain of computational analysis in previous academic works. Our joint group of Zhejiang University and China Academy of Chinese Medical Sciences (CACMS) took the first systematic approach of *Semantic Web for TCM Informatics*, aiming at the computerization and integration of TCM information and knowledge to provide intelligent Web resources for clinical decision-making, drug discovery, and education. The resulting Semantic Web platform deployed in CACMS, integrates over 70 legacy relational databases into a coherent semantic view, providing various Web-based knowledge and information services for TCM practitioners from CACMS's 17 affiliated institutions in China [1].

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

A set of semantic-based tools and systems are developed and deployed to facilitate TCM practitioners in achieving collective intelligence. The Unified TCM Language System (UTCMLS) is the largest TCM Semantic Web ontology including 5,000 concepts and 20,000 instances, serving as a common knowledge representation scheme to improve the quality of semantic search and query, and to infer semantic suggestions such as synonyms and associated concepts. We have also deployed at CACMS the ontology-based query and search engine (Figure 1), which maps legacy relational databases to the Semantic Web layer for query and search across database boundaries. A new methodology named semantic graph mining is proposed, which uses the semantic graph model to integrate graph mining and ontology reasoning for better analyzing biomedical complex networks. The methodology is implemented in the Spora system (Figure 2), which creates knowledge discovery experiments through the orchestration of semantic graph mining services. As the experimental result, the first global herb-drug interaction network is mapped through semantic integration of legacy relational databases in Traditional Chinese Medicine (TCM) domain, and Spora system is applied on this network to discern interesting patterns such as frequent sub-graphs (Figure 3) and community structures (Figure 4). In the resulting network (Figure 4), most nodes (99.3 %) participate in the largest connected components; a small proportion of herbs emerge as hubs through very active connectivity, and they are also at the centrality of the network (based on pair-wise node distance calculation) and serve to connect local drug communities; and there is also a big drug community that consists many biggest hubs in the network, revealing that drug hubs tend to cluster together in TCM domain.

TCM domain experts are interested in these machine-learned patterns rendered as semantic graphs, and realized with amaze that all herbs are connected through decentralized orchestration of formulae in their hands. They evaluate the platform's major technical features as original and productive in TCM drug usage, discovery, and safety analysis, and evaluate the resulting visualized patterns as reflecting TCM practice and potentially leading to a deeper understanding of TCM underlying mechanisms.

The proposed poster contains two components: one will focus on the semantic graph mining methodology, tools, and systems with their Web-based interfaces (Figure 1,2); the other will focus on the description and medical interpretation of computational analysis and knowledge discovery results, including statistical characteristics, frequent subgraphs, and community structures of TCM networks.

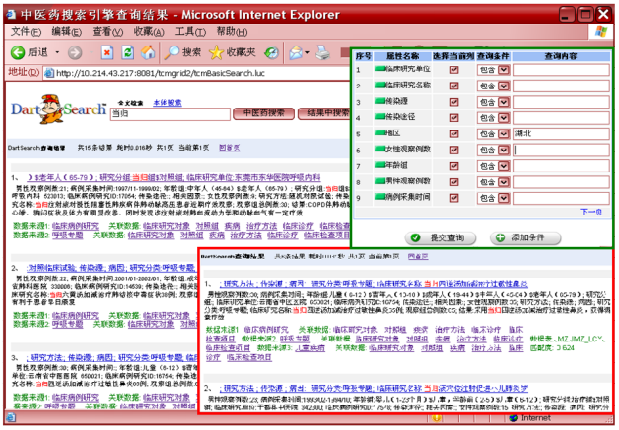

Figure 1: The *DartGrid* semantic query and search portal supports interactive discovery of TCM information and knowledge across database boundaries, through the paradigms of query, search, and navigation based on shared domain ontology.

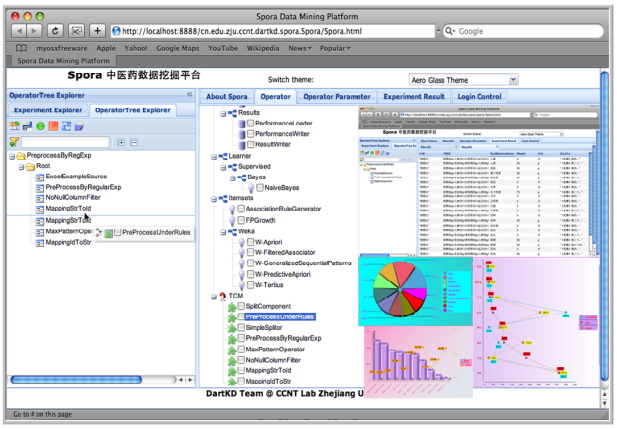

Figure 2: The *Spora* system performs semantic graph mining on the Semantic Web, models a discovery experiment as a tree of operators with customizable properties, and visualizes data mining results through interactive tables, charts, and graphs.

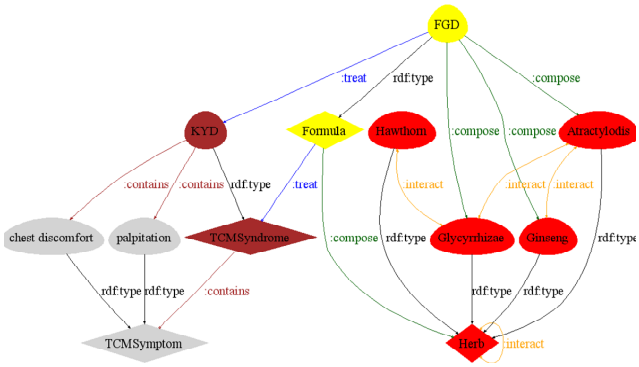

Figure 3: The representation of frequently-occuring patterns as semantic graphs, can help users to understand and interpret these patterns, and make it easier to index, search, merge, and/or compare related patterns. The patterns are obtained by discovering frequently-occuring subgraphs from a set of semantic graphs extracted from heterogeneous data resources, and visualized as a directed, labeled graph for human interpretation.

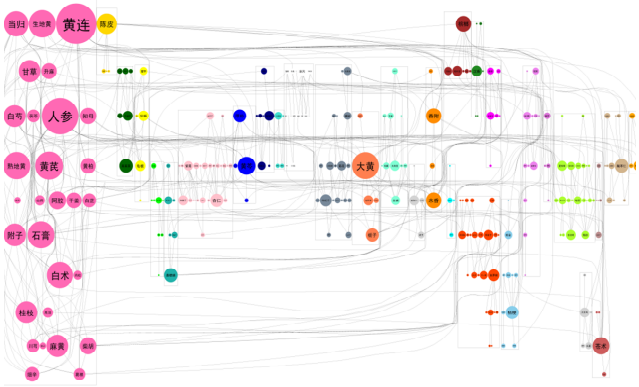

Figure 4: Global network of frequent herb-drug interactions, with drugs represented by nodes with size/font proportional to degree, interactions represented by edges,and drug communities represented by distinct colors. This integrated semantic graph is mapped by merging results of semantic searching for herb-drug interactions from the TCM Semantic Web portal. A clustering algorithm is applied on the graph to decern local communities.

2. CONCLUSION

We presented an in-use Semantic Web platform supporting large-scale database integration, information retrieval, and knowledge discovery for Traditional Chinese Medicine domain. This platform demonstrates the Semantic Web's ability to connect data from interrelated domains for interdisciplinary research, and contributes to the preservation and modernization of TCM as intangible cultural heritage.

3. ACKNOWLEDGMENTS

This work is funded in part by China 973 subprogram NO.2003CB316906, China NSF program NO. NSFC60503018, China 863 program NO. 2006AA01A123, China NSF program No.60525202, and China NSF program No.60533040.

4. ADDITIONAL AUTHORS

Xiaohong Jiang, Yi Feng, Yuxin Mao (Zhejiang University, {jiangxh,fengyi,maoyx}@zju.edu.cn), Heng Wang, Jingming Tang, Chunying Zhou (Zhejiang University, {paulwang,jmtang981,02rjgczcy}@zju.edu.cn).

5. REFERENCES

[1] H. J. Chen, Y. M. Wang, H. Wang, Y. X. Mao, J. M. Tang, C. Y. Zhou, A. N. Yin, and Z. H. Wu. Towards a semantic web of relational databases: A practical semantic toolkit and an in-use case from traditional chinese medicine. In *ISWC 2006: Proceedings of the 5th International Semantic Web Conference*, pages 750–763, Berlin / Heidelberg, 2006. Springer.