Journal of Southeast University (English Edition) Vol. 25, No. 4, pp. 460 – 463 Dec. 2009 ISSN 1003—7985

Intelligent search on integrated knowledge base of traditional Chinese medicine

Fu Zhihong Chen Huajun Yu Tong

(College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China)

Abstract: To semantically integrate heterogeneous resources and provide a unified intelligent access interface, semantic web technology is exploited to publish and interlink machine-understandable resources so that intelligent search can be supported. TCMSearch, a deployed intelligent search engine for traditional Chinese medicine (TCM), is presented. The core of the system is an integrated knowledge base that uses a TCM domain ontology to represent the instances and relationships in TCM. Machine-learning techniques are used to generate semantic annotations for texts and semantic mappings for relational databases, and then a semantic index is constructed for these resources. The major benefit of representing the semantic index in RDF/OWL is to support some powerful reasoning functions, such as class hierarchies and relation inferences. By combining resource integration with reasoning, the knowledge base can support some intelligent search paradigms besides keyword search, such as correlated search, semantic graph navigation and concept recommendation.

Key words: intelligent search; semantic web; knowledge base; semantic index

The world wide web is evolving into a web of data, which contains a set of distributed dataspaces containing heterogeneous resources, such as texts, images, audio/video contents, and structured data. How to integrate these dataspaces to support more intelligent search is a focused theme of the web research community.

One promising solution to the intelligent search is the semantic web, the extension of the current web to incorporate structured data, which provides domain ontologies in languages such as RDF schema and web ontology languages to describe the semantics of information and services. Towards this vision, much progress has been made. A wealth of semantic mapping strategies and tools[1-3] have been developed to make contents of existing legacy databases available for semantic web applications. And still for the large scale of web documents, many automatic or semi-automatic semantic annotation tools such as KIM[4], SemTag[5], MnM[6] and Shoe[7] etc. have been exploited to add formal semantics to the web content. Most of these tools are implemented by the GATE[8] framework. Furthermore, some semantic data integration systems have emerged, such as Dartgrid[9] and RDF based integration[10]. In addition, large amounts of metadata or semantic search engines have been developed to support semantic search with knowledge inference such as Swoogle[11], Falcon-S[12], and OntoSearch[13] etc.

In this paper, we present TCMSearch, an intelligent search engine that is deployed on top of several distributed dataspaces in traditional Chinese medicine (TCM) domain. These dataspaces are maintained by different organizations distributed across China, and manage a large volume of data in various forms, such as relational databases, XML documents, spreadsheets, texts, and other multimedia forms such as pictures. Our goal is to semantically integrate these dataspaces into one virtual database and provide a coherent query interface that provides such functions as instance search, semantic correlation search, and graph-matching.

The core of the system is an integrated knowledge base that uses a TCM domain ontology to integrate heterogeneous resources. The knowledge base is modeled as a graph of instances, in which each data item from relational databases and each document from the web are represented as a resource with a URI. In addition, the semantic relationships between resources are explored so that the data items are correlated. We use machine-learning techniques to generate semantic annotations for texts and semantic mappings for relational databases, so that data items from relational databases and documents are translated into knowledge elements that can be integrated into a knowledge base as a graph of instances. This knowledge base can be enriched as more concepts and semantic relationships are extracted or referred from dataspaces; in turn, a richer knowledge base can be used to discover knowledge from dataspaces more accurately and more deeply.

Within the knowledge base, an ontology based inverted index is constructed and a common access interface in support of instance search and semantic correlated search is provided to retrieve data more appropriately and conveniently. The major benefit of representing the semantic index in RDF/OWL is to support some powerful reasoning functions, such as class hierarchies and relation inferences. By combining resource integration with reasoning, the knowledge base can support some innovative search paradigms besides keyword search, such as correlated search, semantic graph navigation and concept recommendation.

1 System Overview

TCMSearch is an intelligent search engine for traditional Chinese medicine, which integrates several distributed dataspaces semantically and provides some innovative search paradigms such as instance search and semantic correlated search. The architecture of TCMSearch is shown in Fig. 1. It can be divided into three parts: semantic process, index

Received 2009-06-30.

Biographies: Fu Zhihong (1984—), male, graduate; Chen Huajun (corresponding author), male, doctor, associate professor, huajunsir@zju.edu.cn.

Foundation items: Program for Changjiang Scholars and Innovative Research Team in University (No. IRT0652), the National High Technology Research and Development Program of China (863 Program) (No. 2006AA01A123).

Citation: Fu Zhihong, Chen Huajun, Yu Tong. Intelligent search on integrated knowledge base of traditional Chinese medicine [J]. Journal of Southeast University (English Edition), 2009, 25(4): 460 – 463.

process and search process. In the semantic process, data from different sources such as relational databases and web documents are attached with semantic information. For example, relational databases are mapped to correspond ontology-classes according to their table schema; web documents are annotated with ontology instances based on an extensive knowledge base for TCM. In the index process, an ontology-based inverted index is constructed, in which data from TCM dataspace are indexed based on the extracted instances and their content. Finally in the search process, query words are preprocessed semantically and matched in this ontology-based inverted index to search proper results. Furthermore, the search results can be hierarchically navigated by users and semantic correlated search is supported to find relevant information in detail.

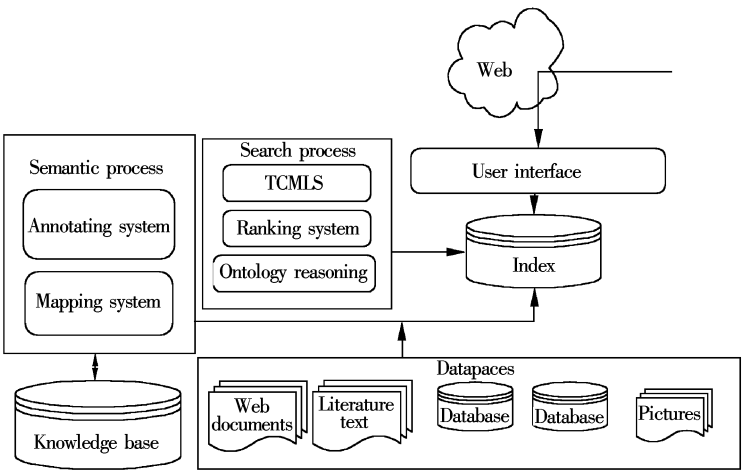

Fig. 1 Architecture of TCMSearch

2 Semantic Model and Representation

The process of modeling TCM dataspace can be divided into three steps. First, there is a prerequisite that a basic knowledge base should be constructed manually, which should at least cover the ontology class system in TCM, and better if there are known instances added at first. Secondly, semantic mapping tools are exploited to bridge the gap between the semantic web and relational databases. Through this method, data records from relational databases can be extracted as instances, and the relationships between them can also be discovered. In turn, this knowledge base can also be enriched by these newly extracted instances and relationships. Finally, semantic annotation for web documents can be automatically done based on the enriched knowledge base, in which not only existing ontology instances and relationships can be extracted but also new ones can be discovered. So the knowledge base can be enriched once again.

2.1 Knowledge base

Knowledge base is the core of the TCMSearch system. It consists of two components, one of which is the ontology class system and the other is a graph of ontology instances and relationships between them. The ontology class system, which is hierarchically organized based on the understanding of TCM and pre-populated by ontology experts manually, defines the schema and guides the process of semantic mapping and annotation. An ontology class defines the type, which represents its features, and the attributes which are categorized into value attributes and object attributes. Generally value attributes are defined by "literal" and used to describe the ontology classes or instances. However, object attributes are defined by other ontology classes and represent the relationships between these ontology classes.

The other component of the knowledge base is much more important and stores real knowledge data, which covers all the instances and relationships between them in TCM. Each instance covers only two of its value attributes, the identifiers of the ontology instance, in TCMSearch system which is "prefLabel" and "alias", and the entire object attributes so that the knowledge base does not grow too huge as new instances and relationships are discovered but ensures the integrity of the instances. In addition, the semantic relationships between instances are explored so that instances in the knowledge base can be interlinked together.

2.2 Modeling structured data

For the purpose of heterogonous relational databases integration, many semantic mapping strategies and tools[1-3] have been put forward recently. All of them follow the schema matching way. The TCMSearch follows the DartMapping[3] tool and is based on the TCM knowledge base.

In this semantic mapping system, a table schema corresponds to an ontology class, of which the columns are mapped to the value attributes and with regards to the joined relationships between tables there are two cases if the joined tables are mapped to the same ontology class. It means that the joined relationship will be used to combine several parts of ontology instances together, so these simple tables will be joined together to represent ontology instances. Otherwise, they should be mapped to the object attributes, in which case it means that instances extracted from these tables are correlated semantically.

Definition 1 An ontology instance is identified by its belonging ontology class "URI" and value attribute "prefLabel" or "alias". And it is described by other value and object attributes.

According to definition 1, data records from relational databases can be modeled as a large graph of ontology instances that are interlinked together by object attributes and class hierarchies. Furthermore, the knowledge base can be greatly enriched by newly coming ontology instances and the relationships between them. It is noticeable that the instances which have the same value for attributing "prefLabel" will only be added just once so that the knowledge base will not be redundant.

2.3 Modeling unstructured data

Nowadays, manual semantic annotation for web documents is very easily accomplished, but it is an expensive process and is often fraught with errors. To overcome the annotation acquisition bottleneck, semi-automatic and automatic annotations of documents have been proposed such as KIM[4], SemTag[5], MnM[6] and Shoe[7] etc. These semantic annotation platforms can be classified into two primary categories, pattern-based and machine-learning-based. In this paper, we combine pattern matching and machine learning methods to do semantic annotation. First, knowledge elements in web documents are discovered and translated to instances according to the instances in the knowledge base. Secondly, the knowledge base checks out the possible relationships between these instances according to the relationships between their belonging classes and hierarchies of the ontology class system. In this way, web documents are annotated by many related ontology instances. Some clustering algorithms are applied to find core instances so that instances for this document are not too large and the semantics is determinable.

3 Ontology-Based Inverted Index

In the traditional information retrieval system, the inverted index technique is always followed. Our ontology-based inverted index is based on it. The documents and database records are not only indexed by their contents but also by their extracted instances. This sort of index structure has two benefits. First, it keeps the nature of the traditional inverted index. Secondly, more appropriate semantic searches can be supported based on the knowledge base, such as instance search to obtain high accurate search results and semantic correlated search to retrieve related information.

3.1 Index implementation

For relational databases, data records are indexed based on the extracted instances. Every database record covers four fields: mapped ontology class, identifiers of the instance, the combination of other value attributes and its database source. The first two fields are for instance search and semantic correlated search. The third field is for traditional keyword search, and the last field is for retrieving original data. However, for web documents, although data records are also indexed based on the extracted instances, their contents are also quite important and can never be discarded. So an index document consists of three parts: annotated instances, ontology classes and the content of the web document. In this way, both instance search and keyword search can be well supported and do not result in loss of information.

3.2 Search support

In the TCMSearch system, keyword search and instance search are both supported. Keyword search is very useful when query words cannot be translated into instances in the knowledge base, and it can make sure that proper search results can be returned only if these query words exist in the inverted index. However, instances search is much more powerful, which can provide more appropriate results and semantic information. As we can see, when query words, after being semantically preprocessed in the TCM knowledge base, can be represented by instances; these instances are searched in the field "identifiers of the instance" of the ontology inverted index and still in the knowledge base so that the search results are very accurate for users. Moreover, the search results can be represented by instances so that semantic correlated search can be supported to conveniently retrieve related information.

4 Conclusion and Future Work

According to the TCMSearch system, dataspaces can be well integrated and instance search can provide more accurate search results by constructing an ontology-base inverted index based on the knowledge base. Moreover, when checking out the details, semantically related results can be queried based on class hierarchies and relation references.

However, there is still much work to be done in the future. First, semantic annotation for web documents is not so good, because so many instances are annotated, although some clustering algorithms have been applied to improve it. Secondly, much deeper ontology inferences should be included in the future. Finally, semantic ranking algorithms should be considered to improve user experience.

References

[1] Zhou Linhua, Chen Huajun, Zhang Yu, et al. A semantic mapping system for bridging the gap between relational database and semantic web [C]//2008 *AAAI Spring Symposium*. Stanford, CA, USA, 2008: 122 − 127.

[2] An Yuan, Topaloglou T. Maintaining semantic mappings between database schemas and ontologies [C]//*Lecture Notes in Computer Science*. Heidelberg: Springer, 2008: 138 − 152.

[3] Chen Huajun, Wang Yiming, Wang Heng, et al. Towards a semantic web of relational databases: a practical semantic Toolkit and an in-use case from traditional Chinese medicine [C] // *Proc of the Fifth International Semantic Web Conference*. Athens, GA, USA, 2006: 750 − 763.

[4] Popov B, Kiryakov A, Ognyanoff D, et al. KIM—a semantic platform for information extraction and retrieval [J]. *Natural Language Engineering*, 2004, **10**(3/4): 375 − 392.

[5] Dill S, Eiron N, Gibson D, et al. SemTag and Seeker: bootstrapping the semantic web via automated semantic annotation [C]//*Proc of the* 12*th International World Wide Web Conference*. New York, USA: ACM Press, 2003: 178 − 186.

[6] Vargas-Vera M, Motta E, Domingue J, Lanzoni M, et al. MnM: ontology driven semi-automatic and automatic support for semantic markup [C]//*Proc of the* 13*th International Conference on Knowledge Engineering and Knowledge Management*. London: Springer-Verlag, 2002: 379 −

391.
[7] Herflin J, Hendler J. Searching the web with SHOE [C]// *Artificial Intelligence for Web Search. Papers from the AAAI Workshop*. AAAI Press, 2000: 35 - 40.
[8] Cunningham H, Maynard D, Bontcheva K, et al. GATE: a framework and graphical development environment for robust NLP tools and applications [C]// *Proc of the* 40*th Anniversary Meeting of the Association for Computational Linguistics*. Philadelphia, USA, 2002: 168 - 175.
[9] Wu Zhaohui, Chen Huajun, Deng Shuiguang, et al. DartGrid: RDF-mediated database integration and process coordination using grid as the platform [C]// *Proc of the* 7*th Asia-Pacific Web Conference*. Shanghai, China, 2005: 351 - 363.
[10] Vdovjak R, Houben G J. RDF based architecture for semantic integration of heterogeneous information sources [C]// *Workshop on Information Integration on the Web*. Eindhoven, The Netherlands, 2001.
[11] Ding L, Finin T, Joshi A, et al. Swoogle: searching for knowledge on the semantic web [C]// *Proc of the* 20*th National Conference on Artificial Intelligence*. AAAI Press, 2005: 1682 - 1683.
[12] Cheng Gong, Qu Yuzhong. Searching semantic web objects with Falcons: approach, implementation and evaluation [J]. *International Journal on Semantic Web and Information Systems*, 2009, **5**(3): 49 - 70.
[13] Thomas E, Alani H, Sleeman D, et al. Searching and ranking ontologies on the semantic web [C]// *Proc of KCAP*05 *Workshop on Ontology Management*: *Searching*, *Selection*, *Ranking and Segmentation*. Banff, Canada, 2005: 57 - 60.

基于中医药集成知识库的智能搜索

付志宏　陈华钧　于　彤

（浙江大学计算机科学与技术学院，杭州 310027）

摘要：为了对异质异构数据资源进行语义集成并提供统一的智能访问接口，利用语义 Web 技术发布机器可理解的数据资源及其之间的关系，以支持智能搜索等功能. 介绍了中医药智能搜索引擎 TCMSearch，该搜索引擎的核心为一个集成语义知识库，该知识库利用领域本体来表示中医药领域的实例及其之间的关系. 首先，针对普通文本，系统采用了机器学习的方法对其进行语义标注；对于关系型数据库数据，则采用了语义映射的方法统一其语义信息. 然后，系统为集成的数据资源构建了一个语义索引，该索引采用本体语言 RDF/OWL 进行表示，从而支持一些强大的推理功能，如类层次关系推理和实例关系推理. 最后，通过利用该语义索引以及其支持的推理功能，系统能够在集成知识库的基础上提供智能化搜索，如关联搜索、语义图浏览以及实例推荐等新功能.

关键词：智能搜索；语义 Web；知识库；语义索引

中图分类号：TP311