Toward Making Online Biological Data Machine Understandable

Cui Tao

Brigham Young University, Provo, Utah 84602, U.S.A.
ctao@cs.byu.edu

Abstract. Huge amounts of biological data are available online. To obtain needed information, biologists sometimes have to traverse different Web sources and combine their data manually. We introduce a system that can automatically interpret the structures of heterogeneous Web pages, extract useful information from them, and also transform them to machine-understandable pages for the Semantic Web, so that a Semantic Web agent can automatically find the information of interest.

Huge and growing amounts of biological data reside in various online repositories. Most of them only focus on some specific areas or only allow limited types of user queries. Sometimes the information a user needs spans multiple sources. A system that traverses only one source may not answer user queries completely. A system that can automatically retrieve, understand, and extract online biological data independent of the source is needed. In this research, I propose a system that can automatically interpret, extract, and annotate biological data with respect to an ontology and make it machine understandable as Semantic Web data. This research interweaves many different areas in information technology and bioinformatics. In [3], I surveyed different approaches in Information Extraction, Schema Matching, the Semantic Web, Data Integration, and Bioinformatics, in order to prepare for my own research.

The first step is to understand heterogeneous source pages automatically, which means to recognize attribute-value pairs and to map these attribute-value pairs to the concepts in the domain extraction ontology[1]. We proposed two techniques to resolve this problem, *sibling-page comparison* and *Sample-Ontology-Object recognition.* The sibling page comparison technique compares *sibling pages*, which are the pages commonly generated by underlying web databases, and identifies and connects non-varying components as category labels and varying components as

[1] The system works based on a data extraction ontology, which is a conceptual-model instance that serves as a wrapper for a domain of interest [2]. When an extraction ontology is applied to a Web page, the ontology identifies objects and relationships and associates them with named object sets and relationship sets in the ontology's conceptual-model instance and thus wraps the recognized strings on a page and makes them "understandable" in terms of the schema specified in the conceptual-model instance.

I. Cruz et al. (Eds.): ISWC 2006, LNCS 4273, pp. 992–993, 2006.

data values. Experimental results show that it can successfully identify sibling tables, generate structure patterns, and interpret different tables using the generated patterns. An alternate way to discover attribute-value pairs and map them to concepts in the ontology is through the use of a sample ontology object. A *sample ontology object* contains as much information as we can collect for one object in a specified application domain with respect to the extraction ontology. For a sample ontology object to be useful, it must commonly appear in many sites. Instead of attributes, our sample-ontology-object recognition technique depends on values to detect structural patterns and tries to infer the structure pattern of a page by observing the layout of the page with respect to the sample ontology object.

If we can interpret a source page and have already matched attribute value pairs in the source page to target concept(s) in the ontology, it is not hard to semantically annotate values for each page in the site using the ontology as the annotation ontology since the machine has already "understood" it [1]. This means that we can transform a source page to a Semantic Web page, which is machine-understandable.

The system is partially implemented. I have implemented tools to interpret source tables and finished a few papers related to this topic [4]. I am currently working on generating a set of sample ontology objects, implementing the sample-ontology-object recognition technique, and building a tool that can semi-automatically generate ontologies in the molecular biology domain from source tables and a few sample ontology objects. To build an ontology in such a broad domain is not easy, not to mention that it should to be automatic. I am currently facing many challenges such as how to better resolve the scalability issues and inter-sources conflicts; and what kind of information we should cover.

The prototype system is to be built for research purposes. It will not do any integration beyond synchronization with the target extraction ontology. The extraction ontology will not cover all the concepts, relationships, and values in the molecular biology domain. Although I will implement and test the system in the molecular biology domain, this approach will likely be general to all application domains that have similar characteristics.

References

1. Y. Ding, D. W. Embley, and S. W. Liddle. Automatic creation and simplified querying of semantic Web content: An approach based on information-extraction ontologies. In *Proceedings of the 1st Asian Semantic Web Conference (ASWC'06)*, 2006. to Appear.
2. D.W. Embley, D.M. Campbell, Y.S. Jiang, S.W. Liddle, D.W. Lonsdale, Y.-K. Ng, and R.D. Smith. Conceptual-model-based data extraction from multiple-record Web pages. *Data & Knowledge Engineering*, 31(3):227–251, November 1999.
3. C. Tao. Biological data extraction and integration — a research area background study. Technical report, Brigham Young University, UT, USA, May 2005.
4. C. Tao and D. W. Embley. Table intepretation by sibling page comparison. 2006. submitted.