Extracting Ontology Properties from the Web-Tables

Song-il Cha, College of Information Science and Engineering, Northeastern University, Shenyang, China
Z. M. Ma, College of Information Science and Engineering, Northeastern University, Shenyang, China

ABSTRACT

Web-tables are ubiquitous in Web pages. Since tables themselves are organized structurally and semantically, they are good resources from which we can easily extract ontology. But, most Web-tables are designed for intuitive perception of humans, thus, it has a certain limit to interpret table content using only structural information of the table. So this paper focuses on the method for interpretation of table content based on semantic characteristics of the table. In order to obtain many property elements used for ontology inference, in this paper, the authors discuss how to extract ontology properties from Web-tables. The extracted properties include the following elements: Is-a relationship, class-instance relationship, triple, property domain, property range, symmetric property, transitive property, functional property, and inverse functional property, property for defining super-sub relationship. Through experiment, the authors show that their method can effectively extract property elements from Web-tables.

Keywords: Class-Instance Relationship, Is-A Relationship, Ontological Properties, Property Domain, Property Range, Tables, Web Pages, Web-Tables, World Wide Web

1. INTRODUCTION

Recently, the Semantic Web comprises techniques that promise to dramatically improve the current WWW (World Wide Web) and its use. With the emergence of the Semantic Web and the growing number of heterogeneous data sources, the benefits of ontologies are becoming widely accepted. Web ontologies define terms used as data (metadata) for explaining things of a special domain. Nowadays, researchers are paying attention to automatic transformation of Internet resources in the areas into ontologies (Antoniou & van Harmelen, 2004). The Web is an enormous source of information contained in billions of individual pages. Most information resource on the Web is presented in the form of semi-structured or unstructured documents, encoded as a mixture of loosely structured natural language text and template units.

Web-tables are used mainly for structuring information, and they are the important means of presenting structured information. Table structures represent relations between data in the table. Therefore, we can easily extract ontology from a table based on the features of table structures without the use of syntax analysis.

DOI: 10.4018/jssoe.2012070104
(Tanaka & Ishida, 2006). However, understanding of table contents requires table structure comprehension and semantic interpretation, which exceed the complexity of corresponding linguistic tasks. Previous studies for extracting information from Web-Tables are centralized to interpret table structure: Pivk et al., (2004) focused on understanding table-like structures only due to their structural dimension, and the table model consists of Physical, Structural, Functional and Semantic components. The authors transformed the most relevant table types into F-logic frames, and also demonstrated and evaluated the successful generation of frames from HTML (HyperText Mark-up Language) tables. Jung et al., (2006) suggested a method for extracting table-schemata based on table structure and heuristics. Using this method, a table is converted into a table-schema and a triple. Moreover, since it is important to determine whether or not a table has meaningfulness that is related to the structural information provided at the level of the table head, the authors further investigated the types of tables, established the features that distinguished meaningful tables from others, built a training data set using these features and constructed a classification model using a decision tree. However, the work in Jung et al. (2006) cannot handle a table without head and also cannot extract an appropriate head using background color and font. Further, Chavan & Shirgave (2011) introduced a method for determining the meaningfulness of a table and extracting the head contents from meaningful table. The authors applied table mining to general HTML documents, separated meaningful tables from decorative tables, and extracted the information using head.

As specified above, it can be found that the work mentioned above mainly focused on how to extract information from tables, instead of investigating how to extract ontologies from Web-tables, and the detailed ontology extraction rules are not provided. In addition, most of research endeavors to interpret the table by using structural characteristics of the table. But, many Web-tables are designed by humans, thus, it has a certain limit to automatically interpret tables using only structural information of the table. Jung et al., (2007) detected that, generally, a table provides a semantic core element in a head, and proposed a method for obtaining class-instance relationship and triple using heuristics for extracting table schemata based on semantic core element. Although the authors in Jung et al. (2007) proposed heuristics for detecting semantic characteristics based on the location of table cells, they did not mention which becomes semantic core element.

Through the experimental observation about semantic features of the Web-tables, we found that if there are proper nouns on the table, then they can become a semantic core element. The previous research (Cha et al., 2012) proposed algorithms to automatically generate total instances or properties belonging to ontology class from a Web table, and take several instances or properties belonging to the class composed of proper noun as the seed. Also, the previous research (Cha et al., 2011) proposed a method for classifying the Web-tables and extracting several relationships. The work focused on interpreting the property range as string only, and thus it can’t extract many property elements. Based on the observation above, the existing work for extracting table information did not discuss how to extract the sufficient property elements used for ontology inference such as property with hierarchical relationship and class-instance relationship, properties for restricting properties, and other property types.

To this end, this paper focuses on the extracting property elements based on interpreting table contents by using structural and semantic characteristics of the table. That is, in this paper, we propose a method to construct ontologies with sufficient property that is used for ontology inference.

The article is structured as follows. In Section 2 we present an extraction method of properties with hierarchical relationship and
Related Content

An Introduction to Remote Installation Vulnerability in Content Management Systems

Quality Practices for Managing Software Development in Information System
[www.irma-international.org/chapter/quality-practices-managing-software-development/75741/](www.irma-international.org/chapter/quality-practices-managing-software-development/75741/)

Modulation Recognition for Software Defined Radio Signal
[www.irma-international.org/chapter/modulation-recognition-software-defined-radio/66479/](www.irma-international.org/chapter/modulation-recognition-software-defined-radio/66479/)

Automated Ranking of Relaxing Query Results Based on XML Structure and Content Preferences
[www.irma-international.org/article/automated-ranking-relaxing-query-results/55060/](www.irma-international.org/article/automated-ranking-relaxing-query-results/55060/)

Implementing Background Net with Knowware System for Personalized Keyword Support
[www.irma-international.org/article/implementing-background-net-knowware-system/55063/](www.irma-international.org/article/implementing-background-net-knowware-system/55063/)