ABSTRACT

This article presents an automated approach to integrate multiple analogous ontologies extracted from structured web pages into a common ontology. These ontologies from heterogeneous systems exhibit rich diversity in appearances, structures, terminologies and granularities. We design a unified similarity paradigm that can collect the implicit and explicit evidences that exhibit coherences among ontology and instance, semantic and structure, as well as linguistic and syntactic features. The similarity between ontology elements is derived from three aspects such as intension, extension and context, denoted by <INT, EXT, CXT>, where INT and EXT include corresponding weighted contents from their offspring, and CXT is relevant to evidences shown in their ancestors. The similarity in each aspect is calculated by means of their semantic overlapping and syntactic comparability. We develop a top-down matching algorithm based on matching space selection and similarity reuse; the algorithm facilitates less error-prone mappings and lower computational cost.

Keywords: data models; data schema; knowledge models; ontologies; semantic matching

INTRODUCTION

Ontology integration is one of the most important research topics in knowledge integration and knowledge reconciliation. It has been widely applied in knowledge engineering, database and data warehouse, e-commerce as well as semantic Web (Dill, Eiron, Gibson, Gruhl & Guha, 2003; Maedche, 2002). As more and more organizations in industry, academia and government publish their information (such as services and products) on the Web, how to add semantic annotation for machine-access and thus enhance the interoperability among these autonomous, distributed, and semantically heterogeneous information sources becomes more urgent.

Usually, it is feasible to utilize weighted string-matching method to handle the explicit
evidences exhibited by the co-occurred terms in the ontology elements. The largest obstacles to ontology integration come from the implicit evidences hidden in various representations. As illustrated in Figure 1, the implicit evidences between two diverse ontologies could include: (1) Appearance: besides the various formats used to encode the content, such as RDF/OWL, XML DTD and XML schema, a piece of knowledge can be described using different types of descriptions (e.g., enumeration or property restriction). (2) Structure: the involved elements may be organized using diverse structures (such as plain or nested). (3) Terminology: diverse vocabularies may refer to the identical thing or have similar meaning (such as CPU and Processor). (4) Granularity: content may be expressed by concise version using a simple element vs. complex version using a set of combined elements (such as Hard Disk vs. HD and its offspring). These implicit evidences would thus complicate the problems of merging these analogous ontologies.

In order to achieve the common ontology that depicts the universal namespace of analogous ontologies (relevant to the same kind of objects), we develop a pipeline as shown in Figure 2 to understand the data-intensive objects published on the Web and resolve the interoperability among the data from the heterogeneous systems. The pipeline consists of three steps as follows.

1. **Web→Data**: As Web designers usually use a template or dynamic program to publish all analogous objects (e.g., products and services), all Web pages about these objects appear to share the same presentation style and description framework. By identifying the key difference between the irrelevant components and objects in HTML DOM trees, we first detect object region, which contains only the description of objects. From the object region, we then extract information about the individual product, which we called object data under the interpretation of view syntax model. Details of this step are reported in Ye and Chua (2004b).

2. **Data→Model**: From a set of analogous objects, we induce an object model (ontology or schema) using the procedure outlined by Ye and Chua (2004a, 2006). This model contains the condensed syntax and framework information of objects, which is partially equivalent to the schema of underlying data repository. For instance, notebook ontologies consist of two object

![Figure 1. Two models about notebook (laptop)](image-url)
Related Content

Web 2.0 Social Networking Sites
[www.igi-global.com/chapter/web-social-networking-sites/29288?camid=4v1a](www.igi-global.com/chapter/web-social-networking-sites/29288?camid=4v1a)

A Pattern-Based Method for Re-Engineering Non-Ontological Resources into Ontologies
[www.igi-global.com/chapter/pattern-based-method-engineering-non/64016?camid=4v1a](www.igi-global.com/chapter/pattern-based-method-engineering-non/64016?camid=4v1a)

On the Social Shaping of the Semantic Web
[www.igi-global.com/chapter/social-shaping-semantic-web/35736?camid=4v1a](www.igi-global.com/chapter/social-shaping-semantic-web/35736?camid=4v1a)

Template Based Semantic Integration: From Legacy Archaeological Datasets to Linked Data
[www.igi-global.com/article/template-based-semantic-integration/135560?camid=4v1a](www.igi-global.com/article/template-based-semantic-integration/135560?camid=4v1a)