Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 32 (2007) 349–357
Expert Systems with Applications
www.elsevier.com/locate/eswa

Elicitation synergy of extracting conceptual tags and hierarchies in textual document

Yu-Liang Chi *

Department of Management Information System, Chung Yuan Christian University, 200, Chung Pei Rd., Chung Li 32023, Taiwan, ROC

Abstract

This study develops an ontology building process for extracting conceptual tags and hierarchies in textual corpus. Though humans have been creating ontologies for many years, efficient ontology building processes in textual corpus are extremely ad hoc. Several issues have identified including how to recognize terminology in textual document, name concept tags in terminologies, and derive conceptual hierarchies among concepts. The proposed approach is extraction technique combinations to produce ontology prototype for editors. The empirical feedback indicates that elicitation synergy is productive during the early stages of building. Additionally, this elicitation synergy is especially useful for ontology editors who lack reference models of a working domain and who encounter textual corpus as major knowledge sources.

Keywords: Ontology; Ontology engineering; Concept extraction; Cognition extraction

1. Introduction

Ontologies are built to establish a classification or conceptualization in knowledge related disciplines. Ontologies have long been used to express shared human understanding of information. The use of ontologies by information technology is "*a specification of a conceptualization*" that was defined by Gruber (1993). Moreover, a conceptualization is an abstract, simplified world view used for representational purposes. Noy and McGuinness (2001) summarized the reasons for developing ontology as follows: sharing a common understanding regarding information structure among people or agents, enabling reuse of domain knowledge, and clarifying domain assumptions. Various studies present extensive evidence that ontologies are involved in information technology to improve existing Web-based applications (García-Sánchez, Valencia-García, & Martínez-Béjar, 2005; Staab et al., 2000), in addition to document management (Martin & Eklund, 2000; Motta, Shum, & Domingue, 2000), and agent negotiation (Huhns & Singh, 1997; Khedr & Karmouch, 2005). Ontology techniques also enable knowledge, semantics, and intelligent in application systems. The advantages of using ontology include permitting more disciplined knowledge base design and facilitating knowledge sharing and reuse (Fernandez-Breis & Martnez-Bejar, 2000; van Elst & Abecker, 2002).

Wide agreement exists that when trying to apply ontology-based system experts must focus on specific domain problems and provide common understandings of individual concepts. However, challenges exist in eliciting cognition from the real world and thus designing concepts of ontology. Human experts encounter clear and proper ontologies for using information systems. Thus, building ontologies is extremely time-consuming and requires considerable human effort (Sugumaran & Storey, 2002). Ontology building may become increasingly difficult when either systematical categories or predefined taxonomy is unavailable. For example, building ontology in daily events is harder than in biological nature science. Additionally, human experts can be important in creating ontology, but have difficulty in coming up with widely recognized impersonal perspectives. Restated,

* Tel.: +886 3 2655408; fax: +886 3 2655499.
E-mail address: maxchi@cycu.edu.tw

0957-4174/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2005.11.038

ontology building is more of a craft than an engineering task. To extend referenced sources in building ontologies, various studies have suggested retrieving concepts from documents and from the Web (De Bruijn & Martin, 2002; Gillam, Tariq, & Ahmad, 2005). The abundant textual document not only gives knowledge but also provides raw materials in building ontologies.

This study develops an ontology building process for extracting conceptual tags and hierarchies in textual corpus. Since textual corpus is created for expressing semantics to human understanding, a systematic granular evolution of eliciting stages is required. Thus, the objectives of this study include: (1) Recognizing terminologies based on textual corpus of specific domains; (2) Identifying unambiguous tag name of concept based on terminologies; and (3) Discovering the conceptual hierarchies according to an "is-a" relationship among concepts. To achieve these objectives, this study explored existing extraction approaches, surveyed corresponding tools, and made the revisions necessary to achieve each elicitation stage. To objectively evaluate the proposed approaches, 15 experts were invited to assess the usefulness of this investigation. The empirical results illustrate that the synergy approaches used to derive such an elicitation may be particularly useful in ontology editors that deal with textual corpus as major knowledge sources.

2. Ontology engineering and its contents

2.1. Ontology engineering

Ontology building is more of an art than a science. Though various studies and experience exist on developing ontology building models, no standard methodologies, specifications and approaches exist in this field (Corcho, Fernández-López, & Gómez-Pérez, 2003; Kayed & Colomb, 2002). In practice, it is difficult to devise models that are widely accepted by domain experts (Hui & Yu, 2005). Traditionally, ontology creation is characterized based on a set of activities, as well as associated engineering steps. For example, a comprehensive methodology involves identifying purpose and scope, building engineering (capture, coding, and integrating existing ontologies), assessment, and documentation (Uschold & Grueninger, 1996). Another example of building ontologies presented a seven step lifecycle (Noy & McGuinness, 2001). This lifecycle determines the domain and applied scope to define details including classes, properties, facets, and instances. Meanwhile, numerous research groups have also endeavored to integrate a set of different ontologies. The TOVE (TOronto Virtual Enterprise) provides an example (Fox & Gruninger, 1998).

The above approaches for building ontologies remain labor-intensive. Ontology editors, including domain experts and knowledge engineers, create such systems primarily by hand. Recently, ontologies have been applied to create knowledge bases of various areas. Unfortunately, most areas lack systematic classifications and taxonomic structures for ontology creation reference. Much knowledge is stored in textual sources, such as document and Web pages that frequently have an unstructured format (Gillam et al., 2005; Motta et al., 2000). Consequently, ontology editors encounter the analytical and creative challenges of establishing ontologies in unstructured corpus.

2.2. Ontology contents

Regarding ontology contents, the scope of discrimination can be considered in terms of lightweight and heavyweight ontologies, respectively (Corcho et al., 2003). The former includes concepts, properties, and relationships between concepts. The later includes the above, as well as axioms and constraints on lightweight ontologies. Chandrasekaran, Josephson, and Benjamins (1999) concisely descried definitions of the ontology, which includes terms, their definitions, and axioms.

Owing to the material sources of this study being textual corpus, linguistic perspectives should be considered while building ontologies. Several types of refinement of textual sources may include words, keywords, catalogs, glossaries, thesaurus, and so on. Each type provides different functions for information processing, categorization, and linguistic classification. For example, Huhns and Singh (1997) described four classification schemes that provide semantics for messages among agents. These schemes include keywords, thesauri, taxonomies, and ontologies. Lassila and McGuinness (2001) suggested notions or conceptualizations of potential ontology specification. They presented an ontology spectrum that described the details of their specification. The linear spectrum divides into two parts according to scope discrimination, and each part comprises various notion types that can express classification or structure base inside the ontology. The first part contains catalog, glossary, thesauri, and an informal "is-a" relationship. The extent of the first part resembles the extent of a lightweight ontology. Meanwhile, the later part comprises more notion types including formal "is-a" relationship, value restrictions, and logic constraints. The second part is similar to heavyweight ontology.

Based on this analysis, the "concept" is the most fundamental component of ontology. However, extracting proper concepts from the real world can be personal and error prone. A lightweight ontology is a schema like taxonomy which comprises a conceptual system used to model knowledge. Consequently, ontology editors must first construct a conceptual system, after which editors should identify hierarchical structures among concepts.

3. Proposed ontology construction model

Textual sources may support information retrieval by using keywords, while also using semantic rendering or knowledge awareness. For ontology building, discovering conceptual tags and their hierarchies is the fundamental