Chapter 1
Populating Knowledge Based Decision Support Systems

Ignacio García-Manotas
University of Murcia, Spain

Eduardo Lupiani
University of Murcia, Spain

Francisco García-Sánchez
University of Murcia, Spain

Rafael Valencia-García
University of Murcia, Spain

ABSTRACT
Knowledge-based decision support systems (KBDSS) hold up business and organizational decision-making activities on the basis of the knowledge available concerning the domain under question. One of the main problems with knowledge bases is that their construction is a time-consuming task. A number of methodologies have been proposed in the context of the Semantic Web to assist in the development of ontology-based knowledge bases. In this paper, we present a technique for populating knowledge bases from semi-structured text which take advantage of the semantic underpinnings provided by ontologies. This technique has been tested and evaluated in the financial domain.

INTRODUCTION
Knowledge-based decision support systems (KBDSS) (Klein & Methlie, 1995) are a specific kind of computerized information systems that supports business and organizational decision-making activities on the basis of the knowledge available concerning the domain under question. For a KBDSS to be reliable and accurate, it needs to be backed up with a knowledge base that is extensive, complete and consistent. Only then the system can infer new knowledge and properly support the decision-making process. There is no doubt whatsoever that the Web is the biggest and most dynamic information repository in the world. However, two major challenges hamper the gathering of knowledge from the Web: (1) it is hard to distinguish which information sources
are reliable and helpful and which are not; (2) the transition from the unstructured or semi-structured information available on the Web to machine-processable knowledge is non-trivial.

The focus of our research is on this second issue. Today, most websites do not provide semantic information. Without such semantic information and given the ever-increasing size of the Web, the identification and automatic processing of relevant information is becoming increasingly difficult. In recent years, a number of approaches with the purpose of structuring non-structured and semi-structured data sources have appeared. In particular, some approaches try to automatically associate data and semantic notes to the HTML documents (Hsin-Chang, 2009; Tijerino, Embley, Lonsdale, & Nagy, 2003; Wang, Lu, & Zhang, 1997). Others approaches focus on giving structure to semi-structured documents (Seong-Bae et al., 2009). There are also approaches that attempt to automatically create an ontology from unstructured HTML documents (Du, Li, & King, in press; McDowell & Cafarella, 2008). Ontologies (Studer, Benjamins & Fensel, 1998) can be used to structure information. The formal semantics underlying ontology languages enables the automatic processing of the information in ontologies and allows the use of semantic reasoners to infer new knowledge. The Web Ontology Language (OWL) (Bechhofer et al., 2004) is the ontology language recommended by the World Wide Web Consortium (W3C).

A major problem that hampers the effectiveness of current techniques for structuring non-structured and semi-structured documents is that they provide support for a limited set of resources formats. In this paper, we describe a tool capable of analyzing any kind of Web-available semi-structured document and populating an ontology with the relevant content gathered. The tool is based on a scalable architecture that supports the integration of information coming from heterogeneous Web resources and different data formats (pdf, rss, plain text, html, etc.). In order to accomplish this goal, the system transforms the information retrieved from the different formatted documents into a common representation data structure. The information in this shared representation format is then processed to obtain the instances that will populate the underlying domain ontology. Our approach is backed up by a proof-of-concept implementation that has been tested in the financial domain. The prototype provides support for analyzing the structured content (i.e., tables) in HTML documents.

The remainder of this paper is structured as follows. In Section 2, we provide background information on ontologies, knowledge-based decision support systems and ontology population. OPHERA, the tool for populating ontologies from heterogeneous data sources, is described in Section 3. Details on the implementation of the prototype and its application in the financial domain are given in Section 4. Finally, conclusions and future work are put forward in Section 5.

BACKGROUND

Ontologies

In this work, an ontology is seen as “a formal and explicit specification of a shared conceptualisation” (Studer, Benjamins, & Fensel, 1998). Ontologies provide a formal, structured knowledge representation, with the advantage of being reusable and shareable. In our approach, ontologies are used to represent the knowledge extracted from texts, so ontologies are obtained as a result of knowledge extraction processes.

Ontologies provide a common vocabulary for a domain and define—with different levels of formality— the meaning of the terms and the relations between them. Knowledge in ontologies is mainly formalized using five kinds of components: classes, relations, functions, axioms and instances (Gruber, 1993). Classes in the ontology are usually organized into taxonomies. Sometimes, the