Integrating Ontologies and Bayesian Networks in Big Data Analysis

Hadrian Peter
University of the West Indies, Cave Hill, Barbados

Charles Greenidge
University of the West Indies, Cave Hill, Barbados

INTRODUCTION

Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze (Hoffman, 2013; Hogan, 2013; McKendrick, 2012; *McKinsey Global Institute, June 2011*). In the past the typical sources of such data have been large enterprise databases, but recently much of the data have originated from heterogeneous sources such as the Internet, social media, and smartphones. Although the data from these sources have different characteristics (peculiarities), a major problem with this voluminous amount of data is that it is unstructured, semi-structured, sloppily organized, and lacking adequate meta-data (Etzioni et al., 2008). One of our challenges, therefore, is how to convert such data into coherent information which can be efficiently utilized in relevant applications. Other familiar challenges posed by big data are the heterogeneity of the data types (variety) and the rate at which the data arrive (velocity).

In performing the big data analysis we use a two-pronged approach. In the first step we use ontologies (Greenidge & Peter, 2010; Peter & Greenidge, 2011) to address the problems associated with the unstructured/semi-structured nature of the big data. However, even after the data have been transformed into an acceptable form, the volume of data has to be reduced to manageable proportions – hence the need for Bayesian networks (Kenett, 2012; Russell & Norvig, 2010; Seigel & Shim, 2003).

The rest of the chapter is organized as follows. In the next section we provide background information on the two (2) main topics of the chapter – namely, ontologies and Bayesian networks. This is followed by the main section in which we discuss the issues, controversies, and problems involved in extracting data from our heterogeneous external data sources. We suggest the ontology approach as one way of solving the problem, and provide the relevant algorithm. We also discuss how Bayesian networks can be used to effectively and efficiently extract knowledge from the resultant large data sets generated. The remaining sections provide the directions for future research and the conclusion, respectively.

BACKGROUND

Ontologies

Many definitions have been advanced for the term "ontology." Some of the more common ones are: Ontology is a strategy for representing knowledge in a consistent fashion; A description of the types of entities within a given domain and the relationships among them. The Power of ontologies lies in their utility for reasoning by means of software applications.

The motivation for our ontology-based framework is a number of earlier approaches to information extraction on the Web. These include hand-written wrappers whose limitations are ro-

DOI: 10.4018/978-1-4666-5202-6.ch115

bustness and scalability hurdles (Crescenzi et al., 2001; Shen et al., 2008); ontology-based matching of data on the Web (Hassell et al., 2006; Embley et al., 1998; Isaac et al., 2007); Jaccard based measures (Euzenat & Shvaiko, 2007); table extraction issues (Holzinger et al., 2006); constructing the MARSON system for performing mappings between relational schema and an (OWL-based) ontology (Hu & Qu, 2007); ontology modeling system for the identification/extraction of instance data from tabular Web pages (Shchekotykhin et al., 2007); key issues in data retrieval (DR), information retrieval (IR), knowledge representation (KR) and information extraction (IE) (Manning et al., 2008); the twin problems of information overload and search (Lee et al., 2008).

Our chapter focuses on mapping Web data, and external data from other sources, to domain ontologies, allowing several IE issues to be directly addressed. We use a variety of techniques to make sense of the structure and meaning of these data, ultimately providing a match to a domain ontology. In particular the WordNet lexical database (Gomez-Perez et al., 2004; Euzenat & Shvaiko, 2007; Fellbaum, 1998) is used to facilitate some basic matching activities. We also make use of current search engine capability in our ontology mapping process. Allowing search engine inputs (and those from other sources) helps us to align the matching process with data as they exist online, rather than as construed in some selectively crafted catalog which may not be representative of Web data (Schoop et al., 2006). Figure 1 is a simple ontology showing the different types of relationships between entities, and is adopted from (Greenidge & Peter, 2010).

Figure 1. Simplified ontology