Information Processing and Management 52 (2016) 765–782

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

FCA based ontology development for data integration

Gaihua Fu*

School of Civil Engineering and Geosciences, Newcastle University, Newcastle upon Tyne, UK

ARTICLE INFO

Article history:
Received 12 May 2015
Revised 5 November 2015
Accepted 22 February 2016
Available online 14 March 2016

Keywords:
Ontology development
Formal concept analysis
Data integration
Information sharing

ABSTRACT

Data is a valuable asset to our society. Effective use of data can enhance productivity of business and create economic benefit to customers. However with data growing at unprecedented rates, organisations are struggling to take full advantage of available data. One main reason for this is that data is usually originated from disparate sources. This can result in data heterogeneity, and prevent data from being digested easily. Among other techniques developed, ontology based approaches is one promising method for overcoming heterogeneity and improving data interoperability. This paper contributes a formal and semi-automated approach for ontology development based on Formal Concept Analysis (FCA), with the aim to integrate data that exhibits implicit and ambiguous information. A case study has been carried out on several non-trivial industrial datasets, and our experimental results demonstrate that proposed method offers an effective mechanism that enables organisations to interrogate and curate heterogeneous data, and to create the knowledge that meets the need of business.

1. Introduction

Business productivity and competitiveness are increasingly being driven by the effective access and use of data. Data provides a mine of information that can help us spot undiscovered patterns of business importance and to create the knowledge that will be needed to tackle the challenges of future. However with data becoming available and growing at unprecedented rates, organisations struggle to take full advantage of valuable data. One main reason for this is that data is usually created and maintained by a range of organisations. This results in mismatch between datasets, i.e., datasets differ from one organisation to another not only in *what* is encoded but also in *how* it is encoded.

In order for organisations to use and digest heterogeneous data and uncover the untold business patterns, there is a growing interest to develop techniques that investigate complex data phenomena and facilitate better data interoperability (Doan, Halevy, & Ives, 2012; Doan, Noy, & Halevy, 2004; Duckham & Worboys, 2005; Huang, Lin, & Chan, 2012; Jiang, Zhang, Tang, & Nie, 2015; Lenzerini, 2002). Among various techniques developed, ontology research is one discipline that can deal with data heterogeneity and improve data sharing (Kalfoglou & Schorlemmer, 2003; Mate et al., 2015; Noy, 2004). Ontology-based integration systems are usually characterised by a *global* ontology which represents a reconciled, integrated view of the underlying data sources. Systems taking this approach usually provide users with a uniform interface—all queries made to source data are expressed in terms of a global ontology, as are the query results. This frees the user from the need to understand each individual data source. Unfortunately, in many domains one faces the problems of either having no

* Corresponding author. Tel: +441912086822.
E-mail address: Gaihua.fu@ncl.ac.uk, fugaihua@hotmail.co.uk

http://dx.doi.org/10.1016/j.ipm.2016.02.003

established ontology that can be readily employed in the integration work, or existing ontologies do not fit for the purpose (e.g., not consisting of knowledge that sufficiently captures the semantics of the information under investigation).

In this paper we contribute a formal and semi-automated approach for ontology development. Rather than starting from scratch, we build an ontology by effective discovering and use of the knowledge that is buried in the datasets to be integrated. The method is based on Formal Concept Analysis (FCA) (Ganter & Wille, 1999; Ganter, Stumme, & Wille, 2005), which is a mathematical approach for data analysis. FCA supports ontology development by abstracting conceptual structures from attribute-based object descriptions, and it enables considerable ontology development activities automated.

Our research extends classical FCA theory to support ontology development for integrating datasets that exhibit *implicit* and *ambiguous* information. Implicit information is caused by the fact that some organisations tend to take some domain knowledge as granted, and do not explicitly specify it in their design documents or datasets. This can lead to an ontology that is ill-formed, and does not correctly capture critical concepts and the semantics of the domain. Ambiguous information is due to the fact that organisations differ from each other in culture, conventions and requirements in system development, hence they may vary in how they choose to represent a business object, and at what levels of granularity such information is encoded. This causes inconsistencies between the datasets of different organisations.

We consider that overcoming this implicity and ambiguity is an important step in ontology development. The work reported here is a follow on research of Beck et al. (2013), Fu and Cohn (2008a) and Fu and Cohn (2008b). In this paper we report further technical advances we have made. To restore implicit information, we introduce a rule based method. We discuss how rules are derived and deployed for recovering implicit information. To resolve disambiguate information, we define a set of primitive operations to deal with simple matches in data alignment. These operations are then composed to deal with more complicated matches. Finally, we report on our experiments that are carried out to construct an ontology for integrating non-trivial datasets from several UK water companies. We measure the quality of the developed ontology by utilising the metrics of classical information theory and also in terms of its *fitness* to the application domain. Our experimental results demonstrate that techniques described in this paper provide an effective mechanism for reconciling and harmonising heterogeneous data from disparate sources, and they support development of ontologies that better *fit* and *respect* the underlying knowledge structures of domains.

The remaining part of the paper is organised as follows. Section 2 reviews related research. Section 3 recalls relevant notions of FCA and briefs our framework for ontology development. Sections 4 and 5 present techniques that deal with implicit and ambiguous information. Section 6 discusses how to derive an ontology by using results generated from Sections 4 and 5. Section 7 reports our experimental results. Section 8 concludes the paper and suggests future research.

2. Related research

Several areas of research are interesting to this work. Firstly, integration techniques investigated in database and information integration are quite relevant. Various topics have been studied by these communities and the ones that are the most interesting here are *mapping discovery* and *schema integration*, and techniques have been developed to support these (Bahga & Madisetti, 2015; Do & Rahm, 2002; Doan et al., 2004; Lenzerini, 2002; Liu & Zhang, 2014; Madhavan & Halevy, 2003; Pedersen, Pedersen, & Riis, 2013; Rahm & Bernstein, 2001). *Mapping discovery* takes two or more database schemas as input and produces a mapping between elements of the input schemas that correspond semantically to each other. Many of the early as well as current mapping solutions employ hand-crafted rules or heuristics to match schemas (Madhavan, Bernstein, & Rahm, 2001; Rahm & Bernstein, 2001). Examples of such heuristics include linguistic matching of schema element names, detecting similarity of structures of schema elements, and considering the patterns in relationships of the schema elements. Techniques have also been proposed to use learning based methods (Doan, Domingos, & Halevy, 2001; Neumann, Ho, Tian, Haas, & Meggido, 2002).

Schema integration constructs a global schema based on the inter-schema relationships produced in mapping discovery. Each mapping element is analysed to decide which representation of related elements should be included in the global schema. When a mapping describes the corresponding schema elements as identical, their integration is straightforward—simply includes one of schema elements into the global schema. More frequently, the corresponding schema elements are not the same but are mutually related by some semantic properties, and schema merging is performed manually or semi-automatically with the assistance of domain engineers to guide the designers in their resolution.

Ontology research is another discipline that deals with data integration. A common definition of an ontology is that it is a formal, explicit specification of a domain of discourse (Gruber, 1993). As it provides a shared understanding and explicit specification of a domain, an ontology is considered to have a key role to play in data integration (Bakhtouchi, Bellatreche, & Ait-Ameur, 2011; Bian, Zhang, & Peng, 2011; Noy, 2004; Uschold & Grüninger, 2004; Yu et al., 2012). Unfortunately, for many domains one faces the need to develop ontologies from scratch (as there is no existing ontology that can be used readily), and a growing number of methods have been proposed in recent years to address the issues of ontology design and development. Most methods are based on the traditional knowledge engineering approach (Brockmans et al., 2006; Pinto & Martins, 2004; Sure, Tempich, & Vrandecic, 2006). These methods usually start with defining the domain and scope of ontologies. This is followed by a data acquisition process: important concepts are collected; a concept hierarchy is derived, and properties and semantic constraints are attached to concepts.

As developing ontologies from scratch is an expensive process to perform, there has been increasing interest in reusing or merging existing ontologies (or other knowledge structures such as thesauri) that are developed independently in different