Research on Multi-Source Data Integration Based on Ontology and Karma Modeling

Hongyan Yun, College of Computer Science and Technology, Qingdao University, Qingdao, China
Ying He, School of Electronic Information, Qingdao University, Qingdao, China
Li Lin, College of Computer Science and Technology, Qingdao University, Qingdao, China
Xiaohong Wang, Qilu University of Technology, Shandong Academy of Science, Shandong Computer Science Center, Shandong, China

ABSTRACT

The purpose of data integration is that integrates multi-source heterogeneous data. Ontology solves semantic describing of multi-source heterogeneous data. The authors propose a practical approach based on ontology modeling and an information toolkit named Karma modeling for fast data integration, and demonstrate an application example in detail. Armed Conflict Location & Event Data Project (ACLED) is a publicly available conflict event dataset designed for disaggregated conflict analysis and crisis mapping. The authors analyzed the ACLED dataset and domain knowledge to build an Armed Conflict Event ontology, then constructed Karma models to integrate ACLED datasets and publish RDF data. Through SPARQL query to check the correctness of published RDF data. Authors design and developed an ACLED Query System based on Jena API, Canvas JS, and Baidu API, etc. technologies, which provides convenience for governments and researches to analyze regional conflict events and crisis early warning, and it verifies the validity of constructed ontology and the correctness of Karma modeling.

KEYWORDS

Armed Conflict Event Ontology, Armed Conflict Location, Baidu API, Canvasjs, Data Integration Approach, Event Data Query System, Jena API, Karma Modeling, SPARQL Query

INTRODUCTION

Big data is widely described as having three dimensions: volume, velocity, and variety (Knoblock & Szekely, 2015). Volume refers to the problems of how to deal with large amount of data sets. Velocity refers to dealing with real-time streaming data, where it may be impossible to store all data for later processing. Variety refers to dealing with multiple types of sources, and different formats of the data. In real life, people need to process multi-source heterogeneous data frequently. Data integration aims to integrate data from multiple heterogeneous data sources together so that users can ignore semantic differences and structural differences (Noy, 2004).

Ontology provides an effective representation to represent concepts and the semantic relationships among concepts (Senthilnayaki, Venkatalakshmi, &Kannan, 2015). Moreover, ontology can be easily expressed by using formal semantic markup languages such as RDF and OWL. Ontology cannot only effectively solve the problem of multi-source data description, but also can break through the bottleneck of semantic mapping. Using ontology for data integration cannot only deal with structure and semantic differences among datasets, but also can be as the basis of data query and reasoning. This

DOI: 10.4018/IJIIT.2019040105

Copyright © 2019, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
paper focuses on exploiting semantic technology to solve the problem of multi-source heterogeneous data variety.

We propose an approach to integrate data from multiple types of sources (for example, spreadsheets, relational databases, web services, and others) and in widely different formats including both relational and hierarchical data (that is XML or JSON). In this approach, domain ontology is used to describe data sources semantically, and semantic integration for multi-source data is realized by using an information integration tool named Karma (Knoblock & Szekely, 2008) to map multiple datasets to RDF data. The use of semantics in this integration process is key to building an approach that scales to large numbers of heterogeneous sources. In this paper, according to the proposed method framework, authors demonstrate the integration and application of ACLED (Armed Conflict Location & Event Data Project) (Raleigh et al., 2017) data in detail.

ACLED is a publicly available conflict event dataset designed for disaggregated conflict analysis and crisis mapping. This dataset codes the dates and locations of all reported political violence and protest events in over 60 developing countries, including Africa and Asia. The data come from news reports, publications by civil society and human rights organizations, and security updates from international organizations. Information is recorded on the battles, killings, riots, and recruitment activities of rebels, governments, militias, armed groups, and protesters. ACLED has recorded close to 200,000 individual events, with ongoing data collection focused on Africa and ten countries in South and Southeast Asia. These data can be used for immediate and long-term analysis and mapping of political violence and protest across developing countries through use of historical data from 1997, as well as informing humanitarian and development work in crisis and conflict-affected contexts through real time data updates and reports.

The contributions of this paper are four aspects. First, authors propose a data integration approach based on ontology and Karma modeling to fast integrate data from multi-type (for example, spreadsheets, relational database, web service, and others) and multi-format (including both relational and hierarchical data, such as XML or JSON). According to this proposed approach framework, Users can quickly realize the integration of multi-source heterogeneous data in a certain field. Second, authors demonstrate the method of domain analysis and construction of domain ontology. Then construct Armed Conflict Event Ontology to realize data sources description. Third, authors demonstrate Karma modeling to realize semantic map and publish into RDF (W3C Recommendation, 2014) data in detail. SPARQL (Simple Protocol RDF Query Language) (W3C Recommendation, 2013) query is used to check the correctness of published RDF. Published RDF data has canonical syntax and correct semantics to facilitate the use and development of users and service integrator. Fourth, authors present a practical development framework of integrated data query system based on constructed ontology and published RDF data. Authors design and develop an application system named ACLEDQS (ACLED Query System) to implements ACLED data query, statistical analysis, and comparison functions. The application system demonstrates how to reuse RDF data conveniently.

RELATED WORK

The problem of describing semantics of data sources is at the core of data integration (Doan, Halevy, & Ives, 2012) and exchange. In traditional data integration, the mapping generation problem is usually decomposed in a schema-matching phase followed by schema mapping phase. Schema matching finds correspondences between elements of the source and target schemas. Schema mapping defines an appropriate transformation that populates the target schema with data from the sources (Taheriyani, Knoblock, Szekely, & Ambite, 2016). Mappings may be arbitrary procedures, but of greater interest are declarative mappings expressible as queries in SQL, XQuery, or Datalog. These mapping formulas are generated by taking into account the schema matches and schema constraints. Alexe et al. (2011) generate schema mappings from examples of source data tuples and the corresponding tuples over the target schema. They create a graph from the conceptual model (CM) of each schema and then
Related Content

Optimal Strategy for the Smartphone Industry in Taiwan: HTC Case Study
www.igi-global.com/article/optimal-strategy-smartphone-industry-taiwan/74830?camid=4v1a

Heuristic Optimization-Based Clustering Solution for Large Facility Location Problems
Tark Küçükdeniz and akir Esnaf (2016). Intelligent Techniques for Data Analysis in Diverse Settings (pp. 171-191).
www.igi-global.com/chapter/heuristic-optimization-based-clustering-solution-for-large-facility-location-problems/150292?camid=4v1a

Supporting Demand Supply Network Optimization with Petri Nets
www.igi-global.com/article/supporting-demand-supply-network-optimization/2427?camid=4v1a