Computers, Environment and Urban Systems xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

An ontology-based spatial data harmonisation for urban analytics

Yiqun Chen[a,*], Soheil Sabri[a], Abbas Rajabifard[a,b], Muyiwa Elijah Agunbiade[a]

[a] Melbourne School of Engineering, The University of Melbourne, Parkville, VIC 3010, Australia
[b] Department of Infrastructure Engineering, Centre for SDIs and Land Administration(CSDILA), Melbourne School of Engineering, The University of Melbourne, Parkville, VIC 3010, Australia

ARTICLE INFO

Keywords:
Ontology
Spatial data integration
Semantic enrichment
Urban analytics

ABSTRACT

Data heterogeneity is one of the most challenging problems in urban data analytics. When obtained from various providers or custodians, datasets for the same domain themes may dramatically differ in formats due to many reasons such as historical legacies, changing definitions or standards across jurisdictions etc. It hinders urban analysts and researchers from understanding and using these data and makes results comparison and interpretation obscure. Ontology, usually created by domain experts, offers a comprehensive representation of knowledge including concepts, relations and properties in a domain. It defines the real world in abstract and offers a universal and stable schema for data harmonisation. This paper proposes a fast, extensible solution for eliminating data heterogeneity by using ontology. Starting from conceptualising domain knowledge to domain ontology, we discuss a two-level mapping mechanism which bonds the nexus between data and ontology using mapping rules. A semantic translation engine is also introduced to automate the data harmonisation process. A real case - urban density indicators computation - also demonstrates the usability of the proposed framework and the results show strong potentials for applying this method to broader urban analytics application scenarios.

1. Introduction

Over the next three decades, more than half of the world's population is expected to live in cities. While cities occupy about 2% of land mass worldwide, they produce more than 80% of global GDP (Dobbs et al., 2011), which is a large economic footprint. Cities also contribute more than 70% of the world's greenhouse gas emissions which add significantly to severe environmental footprint (UN-HABITAT, 2011). In addition to these are several other challenges associated with urbanisation that impact the quality of life, public services accessibility, housing affordability, and health.

Rapid urbanisation worldwide is known to challenge urban planning and management tasks as it brings treats and opportunities for cities. In a current digital era, data management is considered as the main enabler in urban planning, management, and decision-making. However, urban data is still challenged by its notorious heterogeneity (Psyllidis, Bozzon, Bocconi, & Titos Bolivar, 2015; Rajabifard, Ho, & Sabri, 2016). The new streams of big data have further complicated these issues. Big data is usually composed of volumetric and complex data from various sources (e.g. sensor data, social media, and enterprise data) that need classic decision-making organisations to revise their regulatory frameworks for effective utilisation (Sabri, Rajabifard, Ho, Namazi-Rad, & Pettit, 2015). Urban planners are still struggling to interpret the various dimensions of available urban data; particularly when required to understand and plan for complex urban issues such as high-rise building development and its impact on urban temperature. The main challenge is the ability to effectively source, access and leverage the appropriate data for evidence-based planning and decision making.

Urban development analyses involve multi-disciplinary data gathering and analytics (e.g. buildings, infrastructures, populations, and green spaces). As a result, the multi-disciplinary and multi-scale data challenges of urban analytics make the task unique and complex. In general, each discipline has its own data sources that need to be standardised for interoperability, harmonisation and integration for analysis and modelling, fostering complex planning and decision-making tasks.

There have been several initiatives around the world to address the issues of urban data accessibility and interoperability. Examples are the Australia Urban Research Infrastructure Network (AURIN) (Sinnott et al., 2015), Urban Big Data Centre (UBDC) in the UK (Thakuriah, Dirks, & Keita, 2016), and the University of Chicago's Urban Centre for Computation and Data (UrbanCCD) (Catlett et al., 2014). These initiatives and several other similar platforms provide urban researchers

* Corresponding author.
E-mail addresses: yiqun.c@unimelb.edu.au (Y. Chen), soheil.sabri@unimelb.edu.au (S. Sabri), abbas.r@unimelb.edu.au (A. Rajabifard), muyiwaea@unimelb.edu.au (M.E. Agunbiade).

https://doi.org/10.1016/j.compenvurbsys.2018.06.009
Received 15 February 2018; Received in revised form 28 June 2018; Accepted 29 June 2018
0198-9715/

Please cite this article as: Chen, Y., Computers, Environment and Urban Systems (2018), https://doi.org/10.1016/j.compenvurbsys.2018.06.009

and decision-makers unique access to thousands of datasets and analytic tools. However, notwithstanding these initiatives, there are still challenges in harnessing data from different sources and the integration of diverse types of data for robust analyses. For instance, urban planning issues such as decision making for housing affordability need data about land use, income, population density, and transport. While these data might be available through existing platforms as aforementioned, automated data integration is not possible due to data heterogeneity. Another limitation is the provenance and extensibility of data used and models developed within these platforms. As an example, Pettit et al. (Pettit, Tanton, & Hunter, 2017) ascertain that the Shift-Share analysis tool developed in AURIN defined names of derived variables and files but does not provide users with information about input data and the method. As such, the lack of provenance will limit the tools' and models' ability to be scalable and extended to other contexts. Consequently, it underscores the needs to apply the different domains knowledge in determining the semantics of data and their ontology across different jurisdictions while engaging with urban analytics, planning, and management (Catlett et al., 2014; Rajabifard et al., 2016; Thakuriah et al., 2016; Villa, Molina, Gomarasca, & Roccatagliata, 2011).

As such, city planning and policy-making that are location-based and evidence-based reportedly suffer from practical analytics and data-driven decision making due to the lack of access to robust spatial platforms and data sharing infrastructures (Kyttä, Broberg, Tzoulas, & Snabb, 2013; Sabri, Rajabifard, Ho, Amirebrahimi, & Bishop, 2016). In addition, current geospatial databases are used for local- or domain-specific analyses. As a result, city planning and urban development monitoring activities are challenged by the lack of integrated spatial planning and management due to the absence of organised and complex spatial data infrastructures. Substantial work has been undertaken in the past decade. For example, Benslimane et al. (Benslimane, Leclercq, Savonnet, Terrasse, & Yetongnon, 2000) define a spatial ontology to describe key features of urban applications, providing a foundation for semantic reconciliation among heterogeneous spatial information sources. Fonseca et al. (Fonseca, Egenhofer, Davis Jr, & Borges, 2000) propose a creation of software components from diverse ontologies using an object-oriented mapping as a way to share knowledge and data. Raskin and Pan (Raskin & Pan, 2005) develop a collection of ontologies using the web ontology language (OWL) that include both orthogonal concepts (space, time, Earth realms, physical quantities, etc.) and integrative science knowledge concepts (phenomena, events, etc.) for their environmental research. Konstantinou et al. (Konstantinou, Spanos, & Mitrou, 2008) also raise and discuss the problem of mapping relational database contents and ontologies and argue that the addition of formal semantics to the databases is important to make information searchable, accessible and retrievable. Consistent with these efforts, Buccella et al. (Buccella et al., 2011) design and implement a system called GeoMergeP to build a global normalised ontology for integrating geographic data sources. They devise two steps for this purpose. First, by applying a semantic enrichment process on data, a top-level and domain ontology based on the domain ontology of the source and the ISO standards is derived; then continue with a merging process, a shared vocabulary or global ontology is created out of the enriched ontologies. When all data sources are mapped to the global ontology, a federated database is formed for use. Pileggi and Hunter (Pileggi & Hunter, 2017) introduce their ontological approach for establishing the interoperability among heterogeneous datasets for urban indicators computation. In evaluating these previous efforts, a key observation is that most of them parse datasets into a semantic format (e.g., tuples) and provides data discovery and reasoning capabilities by adopting semantic technologies. There are, however, two main drawbacks to these methods. First, by converting and storing datasets as semantic format, an extra copy of data has to be maintained, and it will become intractable for data update and synchronisation. Second, in geospatial and urban analysis domain, a lot of existing models (e.g., road network connectivity, spatial association, agglomeration, clustering, isochrone (Day, Chen, Ellis, & Roberts, 2016; Day, Chen, Ellis, & Roberts, 2017; Yiqun Chen & Rajabifard, 2017; Yiqun Chen, Rajabifard, Spring, Gouldbourn, & Griffin, 2016) and procedures (e.g., spatial union, join, buffer, intersect, clip) are not designed for consuming semantic data format or compatible with semantic technologies. They expect inputs described in traditional geospatial formats while eliminating the heterogeneous data issues, thus, improving their usability.

This paper proposes an ontology-based framework for data heterogeneity elimination by focusing on data accessibility and integration, including provenance and extensibility. It starts with conceptualising domain knowledge and developing this into a domain ontology. It continues with the introduction of a two-level mapping mechanism, which bonds the nexus between data and ontology using semantic enrichment rules. This approach is different from existing methodologies for semantic enrichment of geospatial data, which converts the raw data layer format into a uniformed structure described by the ontology schema. This approach, as explained in Section 4, will mitigate the issue of physically storing any extra data. Section 4 also introduces a semantic translation engine that automates data harmonisation processes. Section 6 explains and demonstrates the usability of the proposed framework in a real case – urban density indicator computation. The last section gives an account of how the proposed framework enables robust urban analytics and decision making and offers suggestions for improvement while speculating on the future directions of ontology-based spatial data harmonisation and urban analytics.

2. From domain knowledge to domain ontology

Ontologies are used for different purposes including intelligent integration of information, the Semantic Web, natural language processing, and knowledge management. From a computer science point of view and in the context of knowledge acquisition, an ontology could be defined as "a formal, explicit specification of a shared conceptualisation" (Staab et al., 2009). In this definition, "formal" refers to the language that is used for the description of the ontology specifications. This language has formal syntax and semantics and makes the ontology descriptions machine readable and machine interoperable. In addition, the word "explicit" refers to all the elements of an ontology, which are explicitly defined (Staab et al., 2009).

In the area of knowledge management (KM), ontology plays a crucial role (Sure, Staab, & Studer, 2009). There are several advantages in developing ontologies for urban analytics. The major advantage is that ontology represents the domain knowledge based on a formal conceptualisation. This type of representation allows the understanding of objects, concepts, and other entities that are assumed to exist in an area of interest (e.g. housing, transport, population) and the relationships that hold among them (Guarino, Oberle, & Staab, 2009). For instance, as illustrated in Fig. 1, the upper level indicates several features in urban density domain, such as urban density parameters, their relationships, and the geographical boundaries. These entities can be conceptualised and structured in a diagram form, shown in the lower level, as a basis for developing urban density ontology.

Another positive side of ontology is the standard way of communication between different domains of knowledge. For instance, an environmental planner might be interested in measuring urban heat island, which deals with the concepts of building height, vegetation, and non-buildable areas. These concepts might be extractable from land use databases with other names such as the number of building stories, green spaces, and open spaces. Hence, the lack of standardised and common language may lead to overlooking the appropriate data.

The other important aspect of ontology is the ability to facilitate better communication between different domain experts by using a standardised common language. Also, ontology facilitates the communication between the human and the machine. It is possible to transfer domain knowledge to domain ontology, which is a machine-readable