A Model to Harness Heterogeneous Data for Urban Traffic Services

Gang-min LI*, Yi-lin ZHAO and Yong YUE

Department of Computer Science and Software Engineering, Xi’an Jiaotong-Liverpool University, Suzhou, China

*Corresponding author

Keywords: Semantic web, OWL, Smart city.

Abstract. The real-time traffic information provides significant convenience to road users and city traffic managers. However, the causes for traffic congestion are rarely and fully unexplored. This paper presents a model to harness heterogeneous data from different sources, explains the potential reasons of traffic phenomena and predict the future traffic flow. A prototype system has been built, which works with real-time, heterogeneous data stream including basic traffic data, planned road works, dynamically events, inclement weather or unplanned street closures. Open data source and social media are captured to diagnose road congestion and explore the underlying causes of traffic congestions. The test result illustrates the feasibility of use proposed model in urban traffic control.

Introduction

With the development of Big Data Analytic and Knowledge Networking technologies, a series of knowledge-intensive applications has been built to abstract knowledge from its underlay data. Semantic Web and Linked Open Big Data are popular approaches in the context of urban planning and traffic control [1]. Applications fuelled by the extensive availability of machine-readable urban data entail reasoning based on the heterogeneous knowledge fusion [2]. Using integrated urban data to determine the time and cause of occurred congestion is an effective way to support transportation management by predicting traffic conditions and making real-time solutions [3]. However, the biggest challenge is analyzing and predicting city traffic phenomena such as accident and temporary incidents because they change over space and time [4]. It is also infeasible to just absorb sensor data without exploring the semantics and relationships among many variables, which are often represented in many different languages and formats [5].

Lack a kind of universal ontology-based semantic mechanism, previous research generally only focused on the representation of traffic conditions and events. The static sensor data is the major resource for analyzing traffic conditions. Another important information source, user (driver) generated data are ignored. For instance, a traffic jam caused by unexpected car accidents may be spread through social media. Traffic sensors such as loop detectors may identify transportation conditions, but are unable to differentiate between different causes of traffic jams. Therefore, merging the reliable traditional data with first-time user generated and social media content can help to explain the cause of the traffic conditions and easy road users’ stress.

The paper reports a research which aims not only report traffic condition but also explain the cause of the condition. Our approach is building a new semantic reasoning model for heterogeneous data integration, visualization, and analytics to help the city manager better understand the traffic conditions and help road user to a smoother journey. The paper is organized as follows: section 2 represents our semantic reasoning model; section 3 is the architecture of the prototype system; section 4 reports our primary evaluation results and finally section 5 is the conclusion.

Conceptual Semantic Reasoning Model

Ontology is a discipline that deals with semantic heterogeneity, by establishing formal vocabulary to share knowledge among different applications and facilitate knowledge reasoning [6]. Domain ontologies modeling with descriptive contextual ontology provides a full set of concepts that benefits
in semantic aggregation and exploration, so that achieve the accurate reasoning from knowledge stream that lead to an event.

The proposed semantic reasoning model is a full semantic model capable of catch the spatial-temporal, domain specific and trustworthy urban information and integrates semantic heterogeneous data with indigenous knowledge to analyze, explore and predict the traffic conditions. The proposed semantic reasoning model relies on the W3C semantic Web stack (OWL and RDF) for representing and reasoning semantic information of the knowledge base which consists of city data sources, social media data, drivers’ calls-in data, and static data from city event sources.

**Reasoning Model Logic**

In our semantic reasoning model, the basic technology used is semantic reasoning, which aims at interpreting relevant semantic streams through their location-time changes and correlation. The model (Fig.1) is applied in the traffic context to predict traffic conditions in three aspects:

![Figure 1. Reasoning engine architecture.](image)

**Controlling data variety and rate (A1):** a unified model is created to handle data variety (csv, xml, tweets) and velocity (static, stream), and represent these heterogeneous data streams.

**Reasoning on the evolution of streams (A2):** semantic correlation between time and space in a stream is used to explore information evolution and to identify association of various data streams.

**Prediction in consistency and scale (A3):** predictive reasoning is used to rank and predicate streams in a consistency and scalable way.

The other three important components are:

- **A spatial translator** is used for positing location of data sources, measuring distance between geographic data, and retrieving connected roads.
- **A reasoning engine** is used for checking consistency of streams in a time basis, evaluating sub-assumption and satisfiability and using DL \( EL^++ \) to represent core inference tasks.
- **A triple store** is used for storing and extracting the raw data and historical triples. The B+ Trees indexing structures is used in our model to scale better stream updates.

**Conceptual Specification of A1**

With respect to the conversion of raw data to ontology streams (Fig.2), the processing of three main types of raw data is listed as following:
a) **CSV**: CSV format is a large portion of raw data. In order to convert the CSV to RDF format, convenient tools can be used such as OpenRefine, an open source application for data cleanup and format transformation. Our model embedded a plugin RDFextension which is used to convert CSV data to RDF.

b) **XML**: An XML-to-RDF transformation tool called Krextor is used to convert XML based city events. Existing vocabularies extracted from DBpedia are used for representing predefined capacity and category of events and handling comparison of events. All these events ontologies were matched to the traffic speed ontologies streams by their time interval through the proper interval concept in the W3C Time ontology.

c) **Tweets**: text formatted data, needs a Text-to-Geocoding mechanism to extract location-related keywords for easily assigning a geographic coordinate to a text inputs. In our model, the missing semantics is added by Typifier to extract features such as traffic severity and event.

**Conceptual Specification of A2**

In order to measure similarity, correlation, association rules of events and related variables through knowledge facts, a unified semantic representation model is created. The model is designed to represent semantics of data stream by traffic event ontology, encoded in OWL 2 EL, and the DL $\mathcal{EL}^+$ has been introduced. Fig. 3 is an example, which demonstrates a description of the basic static knowledge for modeling road status.

![Figure 3. Static knowledge of modeling basic road data sample.](image)

In our model. Ontology streams are considered as sequences of ontologies, and each piece of ontology shows a snapshot of a stream at a specific time $t$. A sample stream snapshot $O^m_n(t)$ is shown in Fig. 4: the traffic status of a road r1 located at the 7th Avenue updated every 5 minutes through the traffic speed data stream $O^m_n(t)$ at date and time $t_1$ : 2017-10-22 08:03:00. It is important to use ontological representation to capture the time and reasons evolution across data streams, so as to aggregate different data sources temporally.

![Figure 4. Speed ontology stream at a given time.](image)
Conceptual Specification of $A_3$

In the A2 part of our model, some rules may carry inconsistent prediction whose detected facts are different from other future information. In order to ensure consistent prediction, the number of rules can be reduced by according to their support, confidence and applicability in similar context.

Fig. 5 shows how to combine auto-correlation with association rules for finding the most relevant rules among streams. The specific context such as the same time and the same weather of the real-time exogenous stream is identified and execute auto-correlation with historical knowledge network. Then, the rule whose consequent is harmonious in the current specific context; and is consistent with the captured knowledge from external stream will be selected. It will ensure the rules selected can be used in other similar environment.

![Figure 5. Consistent prediction progress.](image)

Prototype System Architecture

The proposed model has been implemented in a prototype system. Its architecture is presented in Fig. 6. The system with several core components is designed for testing the concepts.

![Figure 6. Service orchestration.](image)

a) **Text-to-Geocode**: It is necessary to convert the user-generated inputs either extracted from social media or from Radio Station needs to transformed into geo format before semantic matching. This mechanism is designed to automatically extract location keywords and map them into a geographic coordinate.

b) **Historical Diagnosing**: A process of diagnosing road traffic congestion built based on the semantic web technologies is trained by historical congested road and city events. The purposes
of traffic diagnosis are to match a traffic abnormality with its possible reasons, and to explain abnormal conditions. The designed system uses semantic web technologies to integrate heterogeneous data and analyzing its semantics.

c) **Semantic Matching**: The matching is done based on similarity, which is computed by two aspects: one is semantic description matching of congestion and its causes; another one is the geographic location of events with the road network. With these, real-time events data including the road network and user-generated content from social media are emerged into the historical knowledge stream as potential reasons based on semantic similarity to previous congestions’ event tags and time.

d) **Semantic Interpretation**: All candidate causes are selected and ranked based on a spatial-temporal confidence. The final presented result is the top three candidate causes.

**Test and Evaluation**

A benchmarking datasets “the NYC traffic Stream” are used and encoded according to the real-time traffic speed stream persisted into a CSV file for experiment. This data represents the averaged vehicle speed of each 135 roads updated every 5 minutes from 2017/10/01 to 2017/10/31. The project develops a simple ontology to represent basic data, using Protégé. City Events are then captured through NYC OpenData and nearly 165 events happened in each day. In addition the project enriches and classifies the events data with semantic rules to capture well-organized semantic descriptions. Each event is described on average through 26 RDF triples. The model is also emerged ontologies from Event ontology [36]. Approximately 40 road constructions records of each day are enriched to overall 1220 RDF triples for a month. The $\mathcal{EL}^{++}$ enrichment of data ensures that historical and real-time data stream can be effectively matched. The basic static ontology knowledge representing road information, events, road closures and road weather data is composed of 56 concepts and 30 object properties. Data of past events, road works, weather and road conditions are stored as $1.2 \times 10^6$ RDF triples in Jena TDB.

The evaluation of our proposed model and the prototype system is focused on two aspects: computing scalability and accuracy. It is done in two main approaches: (1) comparing results with a non-semantic approach in stream prediction [7], and (2) analyzing the speed of model reacting to the number of stream sources.

With the above two evaluation criteria, our prototype system was tested and the results shoes that that the system performed well and satisfy our prove concept purpose. However the system failed when the major source of traffic information contains noise and inaccurate data.

**Conclusion**

This paper reports our efforts to solve inaccuracy and unreason the causes of traffic conditions in urban traffic control systems. We proposed a semantic based model to integrate heterogeneous information and identify the most appropriate vocabulary to represent these semantics. OWL EL was used as semantic encoding of city information to implement the semantic expressivity and semantic similarity computation. Using OWL 2 Full to explore more causes for road congestions, and improve the diagnosis processing. Moreover, other stronger rules that are not limited in the SWRL rules can be used to further reduce rules number, so that increasing the scalability and predicting accuracy.

Our primary implementation and testing demonstrated the feasibility of the proposed model and prototype system design. However, the system exposes the problem that the city sensors often transfer noise information, which can disrupt the semantic processing.

**Acknowledgement**

This research was financially supported by Accenture, LLP, in research using multi-agent system (MAS) in Knowledge Graph to abstract business value out of large operational organization data.
References

[1] T. Kindberg, M. Chalmers, E. Paulos, Guest Editors’ Introduction: Urban Computing, IEEE Pervasive Comput. 6 (2007) 18-20.

[2] G. Brewka, S. Ellmuthaler, R. Gonçalves, M. Knorr, J. Leite, J. Pührer, Reactive multi-context systems: Heterogeneous reasoning in dynamic environments, Artif. Intell. 256 (2018) 68-104.

[3] F. Lecue, S. Tallevi-Diallevi, J. Hayes, R. Tucker, V. Bicer, M. Sbodio, P. Tommasi, Smart traffic analytics in the semantic web with STAR-CITY: Scenarios, system and lessons learned in Dublin City, Journal of Web Semantics - Web Semantics: Science, Services and Agents on the World Wide Web. 27-28 (2014) 26-33.

[4] D. J. Peuquet, N. Duan, An event-based spatiotemporal data model (ESTDM) for temporal analysis of geographical data, Int. J. of Geogr. Inf. Syst. 9 (1995) 7-24.

[5] W. Kuhn, A Functional Ontology of Observation and Measurement, in: K. Janowicz, M. Raubal, S. Levashkin (Eds.), GeoSpatial Semantics, GeoS 2009, Lect. Notes Comput. Sci. Springer, Berlin, Heidelberg, 5892 LNCS (2009) 26-43.

[6] N. Noy, Semantic integration: a survey of ontology-based approaches, SIGMOD Rec. 33 (2004) 65-70.

[7] S. Papadimitriou, C. Faloutsos, Streaming Pattern Discovery in Multiple Time-Series Streaming Pattern Discovery in Multiple Time-Series, Int. Conf. Very Large Data Bases, 2005, pp. 697–708.