STAR-CITY: Semantic Traffic Analytics and Reasoning for CITY*

Freddy Lécué, Simone Tallevi-Diotallevi, Jer Hayes, Robert Tucker, Veli Bicer, Marco Sbodio, Pierpaolo Tommasi

IBM Research, Smarter Cities Technology Centre
Damastown Industrial Estate, Dublin, Ireland
{(firstname.lastname)@ie.ibm.com}

Abstract. This paper presents STAR-CITY, a system supporting semantic traffic analytics and reasoning for city. STAR-CITY, which integrates (human and machine-based) sensor data using variety of formats, velocities and volumes, has been designed to provide insight on historical and real-time traffic conditions, all supporting efficient urban planning. Our system demonstrates how the severity of road traffic congestion can be smoothly analyzed, diagnosed, explored and predicted using semantic web technologies. Our prototype of semantics-aware traffic analytics and reasoning, experimented in Dublin City Ireland and Bologna City Italy, works and scales efficiently with real, historical together with live and heterogeneous stream data.

1 Introduction

As the number of vehicles on the road steadily increases and the expansion of roadways remains static, congestion in cities became one of the major transportation issues in most industrial countries [1]. Urban traffic costs 5.5 billion hours of travel delay and 2.9 billion gallons of wasted fuel in the USA only, all at the price of $\$121$ billion. Even worse, the costs of extra time and wasted fuel has quintupled over the past 30 years. It also used to (i) stress and frustrate motorists, encouraging road rage and reducing health of motorists [2], and (ii) interfere with the passage of emergency vehicles traveling to destinations where they are urgently needed. All are examples of negative effects of congestion in cities.

STAR-CITY (**S**emantic **T**raffic **A**nalytics and **R**easoning for **CITY**), as a system which integrates heterogeneous data in terms of format variety (structured and unstructured data), velocity (static and stream data) and volume (large amount of historical data), has been mainly designed to provide insight on historical and real-time traffic conditions. Most of the existing modern traffic systems such as TrafficView[1] mainly focus on monitoring traffic status in cities using dedicated sensors (e.g., loop indiction detectors), all exposing numerical data. Basic in-depth but semantics-less state-of-the-art analytics are employed, limiting scalable real-time data integration. In such a context semantic expressivity together with reusability of the underlying data is quite limited.

* The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement ID 318201 (SIMPLI-CITY).

[1] https://trafficview.org/

On the contrary STAR-CITY strongly relies on interpreting the semantics of contextual information for deriving innovative insights i.e., analysis, diagnosis [3, 4], contextual exploration [5], and more accurate traffic condition forecasting using recent research work in semantic predictive reasoning [6]. Table 1 reports all data sources processed by STAR-CITY in the Dublin scenario with respect to their velocity i.e., static, quasi stream, stream. They report various types of information coming from static or dynamic sensors, exposed as open, public data and described along heterogeneous formats. All rows in grey are data sets not used for traffic diagnosis [3], but required for prediction.

Type	Sensing	Data Source	Description	Format	Temporal Frequency (s)	Size per day (GBytes)	Data Provider (all open data)
Stream Data		Journey times across Dublin City (47 routes)	Dublin Traffic Department's TRIPS system	CSV	60	0.1	Dublin City Council via dublinked.ie
	Static	Road Weather Condition (11 stations)		CSV	600	0.1	NRA
		Real-time Weather Information (19 stations)		CSV	[5, 600] (depending on stations)	[0.050, 1.5] (depending on stations)	Wunderground
	Dynamic	Dublin Bus Stream	Vehicle activity (GPS location, line number, delay, stop flag)	SIRI: XML-based	20	4-6	Dublin City Council via dublinked.ie
		Social-Media Related Feeds	Reputable sources of road traffic conditions in Dublin City	Tweets	600	0.001 (approx. 150 tweets per day)	LiveDrive Aaroadwatch GardaTraffic
Quasi Stream	Dynamic	Road Works and Maintenance		PDF	Updated once a week	0.001	Dublin City Council
		Events in Dublin City	Planned events with small attendance	XML	Updated once a day	0.001	Eventbrite
			Planned events with large attendance			0.05	Eventful
Static	Static	Dublin City Map (listing of type, junctions, GPS coordinate)		ESRI SHAPE	No	0.1	Open StreetMap

Table 1: (Raw) Data Sources for Dublin City Traffic Scenario.

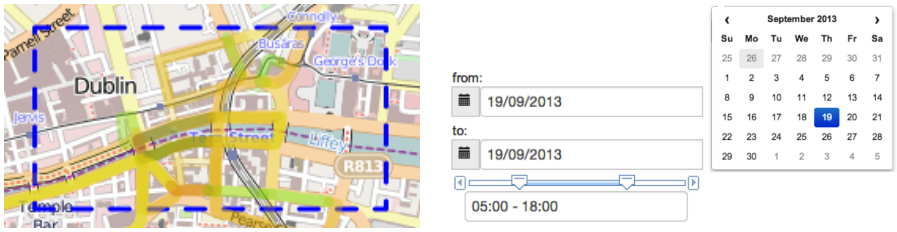
The novelty of STAR-CITY lies in the ability of the system to ingest highly heterogeneous real-time data and perform various types of inferences i.e., analysis, diagnosis, exploration and prediction. These inferences are all elaborated through a combination of various types of reasoning i.e., (i) Description Logic (DL) $\mathcal{EL}^{++}$-based i.e., distributed ontology classification-based subsumption [7], (ii) rules-based i.e., pattern association [6], (iii) machine learning-based i.e., entities search [5] and (iv) stream-based i.e., correlation [6], inconsistency checking [4]. STAR-CITY completely relies on the W3C semantic web stack for representing semantics of information and delivering inference outcomes. Currently applied in the context of Dublin and Bologna Cities, STAR-CITY can scale to any other city, which exposes data sensors of any kind.

This paper is organized as follows. Section 2 sketches one potential scenario using STAR-CITY. Section 3 presents the architecture and functionalities of our system. Finally Section 4 draws some conclusions.

2 STAR-CITY Scenario

The STAR-CITY system is illustrated through a list of scenarios, where each highlights actions that any city manager is required to perform on a daily basis. Its scenarios have been defined with the Dublin and Bologna city transportation department to support actions which are not easily supported by state-of-the-art systems in place (due to the

complexity of data integration and contextual reasoning). The use of semantic web technologies in all scenarios is transparent to end-users. However such technologies are strongly required to compile and deliver contextual analysis, diagnosis, exploration and prediction. All user interactions (UI) are achieved through simple UI paradigms e.g., spatial and temporal selection for initialization (Fig.1.(a) and Fig.1.(b)). All results, delivered by analysis, diagnosis, exploration and prediction, are dynamically exported as parallel, spider, pie, graph-based and time-series charts.



(a) Spatial (Map Area) Selection (b) Temporal (Date / Time Window) Selection

Fig. 1: STAR-CITY Spatio-Temporal Initialization (color print)

For each scenario, we sketch its (i) description, (ii) motivation, (iii) challenge, together with (iv) the STAR-CITY approach, its (v) scalability and (vi) limitation.

2.1 Spatio-Temporal Analysis of Traffic Status

City traffic managers are interested in both historical and real-time information of traffic status (discretized as free, low, moderate, heavy, stopped flow) in order to visually extract the pulse of the city traffic at any time and space.

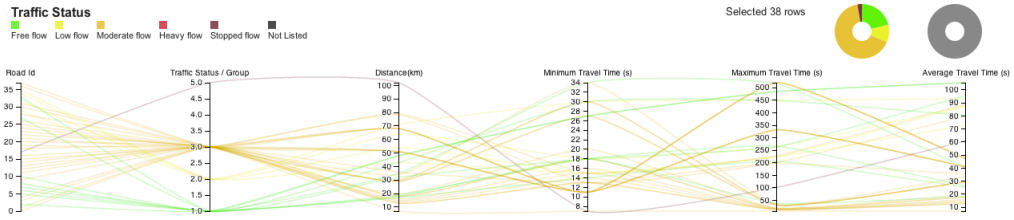


Fig. 2: Spatio-Temporal Analysis of Traffic Status (color print).

In a context of real-time information, stream journey times data needs to be processed in real-time while fast aggregation (average, max, min) is required for historical analysis of traffic status. In both contexts rules-based mechanisms are required to capture and infer traffic status. The STAR-CITY approach consists in discretizing numerical values of travel time individuals (described through road, link, direction, sensors) in status through SWRL[2] rules (OWL $\mathcal{EL}^{++}$ ontologies and associated rules available[3]). Fig.2 embeds the results in a parallel chart, where the status of each road segment together with its proportion (pie chart on the right hand corner) are established. While the

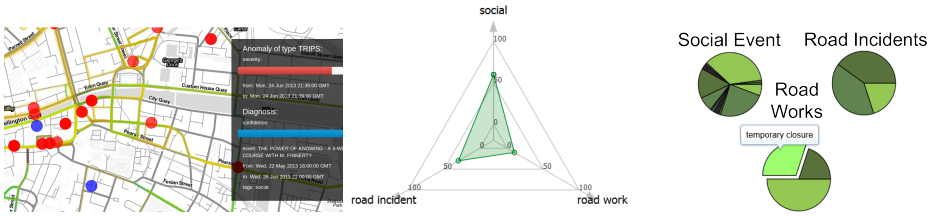
[2] http://www.w3.org/Submission/SWRL/

[3] https://ibm.biz/BdDZ5J

approach is scalable for real-time status under moderate temporal intervals (up to ten weeks), the search and aggregation over tens of months become more challenging.

2.2 Spatio-Temporal Diagnosis of Traffic Status

How to identify the nature and cause of traffic congestion in real-time? How to capture diagnosis results on a spatial and (historical) temporal basis? How to understand the impact of city events on traffic conditions? These are general questions which cannot be answered by existing state-of-the-art traffic systems, but of really importance for city managers to better understand and plan her/his cities at any time. Such question remains open because (i) relevant data sets (e.g., road works, city events), (ii) their correlation (e.g., road works and city events connected to the same city area) and (iii) historical traffic conditions (e.g., road works and congestion in Canal street on July 24th, 2010) are not fully open and jointly exploited. STAR-CITY exploits the DL-based semantics of streams to tackle these challenges. Based on an analysis of stream behavior through change and inconsistency over DL axioms, we tackled change diagnosis by determining and constructing a comprehensive view on potential causes of changes [4]. Some extensions of the latter work have been achieved to support both scalable real-time and historical aggregation of diagnosis results. In addition to a spatial representation of traffic conditions and their diagnosis (Fig.3.(a)), STAR-CITY exposes a spider chart of congestion diagnosis (Fig.3.(b)), and a more in-depth analysis of all causes (Fig.3.(c)), both for any spatio-temporal constraint. Since the diagnosis reasoning of STAR-CITY strongly relies on classification of OWL 2 EL ontologies, we adopt a distributed classification [7] of OWL 2 EL journey times individuals to obtain a scalable diagnosis. The current implementation is limited to $\mathcal{EL}^{++}$ expressivity for scalability reasons.



(a) Diagnosis Interpretation (b) Diagnosis Analysis (c) In Depth Diagnosis

Fig. 3: Spatio-Temporal Historical and Real-Time Diagnosis in STAR-CITY (color print)

2.3 Spatio-Temporal Exploration of Contextual Information

The STAR-CITY system enables the city traffic managers to explore contextual information related to some city events and traffic conditions over historical semantic city data. It helps in terms of giving more insights about the city by retrieving the relevant information from the large city data to find (semantically) similar city events and their impact on previous traffic conditions. Retrieving the relevant contextual information over the heterogeneous and vast city data is a challenging task since classical search techniques are limited in terms of (i) identifying the information needs of the city managers, (ii) handling the contextual information to find similar settings happened in the past, and (iii) utilizing the heterogeneous and semantic data to retrieve accurate information. STAR-CITY addresses these issues following semantic search technologies [8,

5] and extends them significantly to handle both the context and spatio-temporal dimensions. By capturing the current context from the system (spatial, temporal, events, traffic conditions), the system formulates a contextual semantic query which better identifies the actual information need of the city managers within a certain traffic status and city setting. Then, it retrieves the relevant information (e.g., events, traffic conditions) that occurred in a similar context by using its underlying semantic search engine and displays the search results in an exploration interface. This gives the city traffic managers, for example, to get more insight about the similar events in Canal street (or of close proximity) in a similar time of the year and their profound effect on traffic conditions.

2.4 Traffic Status Prediction

Prediction, or the problem of estimating future observations given some historical information, is an important inference task required by city traffic managers for obtaining insight on cities. On the one hand it determines the future states of roads segments, which will support transportation departments and their managers to proactively manage the traffic before congestion is reached e.g., changing traffic light strategy. While predictive analytics spans many research fields, from Statistics, Signal Processing to Database and Artificial Intelligence [9], all existing approaches have been mainly designed for very fast processing and mining of (syntactic and numerical) raw data from sensors. However they rarely utilize exogenous sources of information for adjusting estimated prediction. Inclement weather condition, a concert event, a car accident, peak hours are examples of external factors that strongly impact traffic flow and congestion [10]. They also all fail in using and interpreting underlying semantics of data, making prediction not as accurate and consistent as it could be, specially when data streams are characterized by texts or sudden changes over time. STAR-CITY shows that the integration of numerous sensors, which expose heterogenous, exogenous and raw data streams such as weather information, road works, city events or incidents is a way forward to improve accuracy and consistency of traffic congestion prediction [6]. Fig.4 illustrates how predictions are handled in STAR-CITY. The future status of road segments (in the selected boundary box) and their proportion are reported up to two hours ahead.

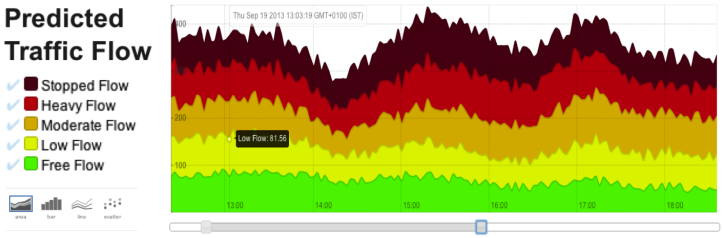


Fig. 4: Prediction of Traffic Status (color print).

Similarly to diagnosis, the scalability of predictive reasoning is highly coupled with the polynomial-time characteristics of subsumption-based reasoning in DL $\mathcal{EL}^{++}$. In the real world, sensors exhibit noise. The causes range from malfunctioning, miscalibration, to network issues and attrition breakdown. Noisy data needs to be detected early to avoid unnecessary semantic enrichment, which could lead to more important problems at reasoning time. We partially addressed this problem by integrating some

custom filter operators at stream processing level to check validity of data. The integration of new data stream needs a careful analysis of historical data in order to identify the most appropriate filters, avoiding as much noise as possible.

3 STAR-CITY Architecture and Technology

This section describes the main technologies behind STAR-CITY.

Semantic Representation: The model we consider to represent static background knowledge and semantics of data stream is provided by an ontology, encoded in OWL 2 EL[4]. The selection of the W3C standard OWL 2 EL profile has been guided by (i) the expressivity which was required to model semantics of data in Table 1, (ii) the scalability of the underlying basic reasoning mechanisms we needed.

Semantic Enrichment: All raw data streams in Table 1 are served as real-time OWL 2 EL ontology streams by using IBM InfoSphere Streams. Different mapping techniques are used depending on the data format. All the ontology streams have the same static background knowledge to capture time (W3C Time Ontology[5]), space (W3C Geo Ontology[6]) but differ only in some domain-related vocabularies e.g., traffic flow type, weather phenomenon, event type. These ontologies have been mainly used for enriching raw data, facilitating its integration, comparison, and matching. The DBpedia vocabulary has been used for cross-referencing entities.

Distributed Semantic Reasoning: Completion $\mathcal{EL}^{++}$ rules [11], used for classification, are distributed across various nodes based on their types. Each node is dedicated to at most one type of (normal form) axioms and runs the appropriate rule on axioms.

Semantic Stream Reasoning: Real-time semantic comparison and matching of stream snapshots are operated. Such computing is required by predictive reasoning and real-time diagnosis for elaborating semantic context (events, weather, incidents) similarity and correlation over time, all in real-time. The stream correlation is established by comparing the number of changes i.e., *new*, *obsolete*, *invariant* ABox entailments between snapshots. The latter ensures context-aware diagnosis and prediction.

Semantic Rule Association and Mining: Predictive reasoning is achieved following state-of-the-art principles i.e., rules association mining. The generation of association rules between streams (and their snapshots) is based on a DL extension of Apriori [12], aiming at supporting subsumption for determining association rules. Contrary to the initial version of Apriori, the association is achieved between any ABox elements together with their entailments (e.g., all congested roads, weather, works, incidents, city events). Rules are encoded in SWRL, and all consequents of each rule are validated though consistency checking. This ensures to obtain consistent, accurate prediction results.

REST Interface : All functionalities of STAR-CITY are exposed through REST services, providing component-ization, evolve-ability via loose coupling and hypertext.

[4] `http://www.w3.org/TR/owl2-profiles/`
[5] `http://www.w3.org/TR/owl-time/`
[6] `http://www.w3.org/2003/01/geo/`

Web User Interface : STAR-CITY strongly relies on HTML, CSS, Javascript (Dojo toolkit, D3, JQuery libraries) to produce an appealing user interface. Time-series, spider charts together with parallel charts are examples where Dojo and D3 components were combined with HTML and CSS.

Deployment : Our technology stack is based on (i) well-established commercial components from IBM e.g., IBM InfoSphere Streams for stream enrichment and processing, IBM WebSphere as the HTTP/Application Server, and (ii) state-of-the-art components such as pssh for parallel distribution of reasoning, Jena TDB[7] as RDF store. We preferred the B+ Trees indexing structures which scale better in our context of many (stream) updates. An alpha version of STAR-CITY, intended for demonstration purposes only, is located at http://www.dublinked.ie/sandbox/star-city/.

4 Conclusion

We presented STAR-CITY, a system which has been designed for (i) aggregating heterogeneous real-time data and (ii) delivering contextual analysis, diagnosis, exploration and prediction of traffic conditions in Dublin and Bologna cities, while being scalable to any city and contexts that involve sensor data. Semantic web technology stack has been deeply used for describing, integrating and reasoning over heterogeneous city data.

References

1. Schrank, D., Eisele, B., Lomax, T.: 2012 urban mobility report. `http://goo.gl/Ke2xU` (2012)
2. Lajunen, T., Parker, D., Summala, H.: Does traffic congestion increase driver aggression? Transportation Research Part F: Traffic Psychology and Behaviour **2**(4) (1999) 225–236
3. Lécué, F., Schumann, A., Sbodio, M.L.: Applying semantic web technologies for diagnosing road traffic congestions. In: International Semantic Web Conference (2). (2012) 114–130
4. Lécué, F.: Diagnosing changes in an ontology stream. In: AAAI. (2012)
5. Bicer, V., Tran, T., Abecker, A., Nedkov, R.: Koios: Utilizing semantic search for easy-access and visualization of structured environmental data. In: ISWC (2). (2011) 1–16
6. Lecue, F., Pan, J.Z.: Predicting knowledge in an ontology stream. In: IJCAI. (2013) ?–?
7. Mutharaju, R.: Very large scale owl reasoning through distributed computation. In: International Semantic Web Conference (2). (2012) 407–414
8. Bicer, V., Tran, T., Nedkov, R.: Ranking support for keyword search on structured data using relevance models. In: CIKM. (2011) 1669–1678
9. Wang, H., Fan, W., Yu, P.S., Han, J.: Mining concept-drifting data streams using ensemble classifiers. In: KDD. (2003) 226–235
10. Cairns, S., Hass-Klau, C., Goodwin, P.: Traffic impact of highway capacity reductions: Assessment of the evidence. Landor Publishing (1998)
11. Baader, F., Brandt, S., Lutz, C.: Pushing the el envelope. In: IJCAI. (2005) 364–369
12. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: VLDB. (1994) 487–499
13. Daly, E.M., Lécué, F., Bicer, V.: Westland row why so slow?: fusing social media and linked data sources for understanding real-time traffic conditions. In: IUI. (2013) 203–212

Appendix - *Tables 2 and 3 summarize how STAR-CITY addresses the SWC requirements*

[7] `http://jena.apache.org/documentation/tdb/index.html`

Criterion	Rating	Explanation
End-User Application	High	Web-based application devoted for ANY user (no computer science / semantic web skills required) interested in city traffic-related matters.
Diverse ownership or control of sources	High	Data is retrieved from various open data sets e.g., Dublin City Council,Dublinked, Wunderground, Twitter ... (see Table 1).
Heterogenous sources	High	SHAPE, tweets, CSV, PDF, XML, RDF, OWL type of format (c.f., Table 1) - all with various type of volume and velocity (for streams).
Real-world data	High	One year of historical data + real-time collection of stream data still ongoing.
Use of Semantic Web Technologies	High	Data is represented in OWL 2 EL and described by domain ontologies: W3C TIME,[5] SPACE,[6] NASA SWEET,[a] IBM TravelTime, IBM SIRI-BUS. Reasoning is achieved through $\mathcal{EL}^{++}$ completion rules.
Data processed to derive useful information	High	City traffic analysis, diagnosis, exploration and prediction.
Suitability of Semantic information processing	High	Context-aware integration is required for achieving accurate and consistent insight. Advantages of using semantic web technologies vs. standard approaches (better accuracy, context-aware reasoning) are described in [4, 3, 6].

[a] http://sweet.jpl.nasa.gov/ontology/

Table 2: Minimal Semantic Web Challenge Criteria Meet STAR-CITY.

Criterion	Rating	Explanation
Attractive and functional Web interface	High	Extensive use of HTML, Dojo, D3, JQuery and geo-mapping tools to produce an attractive interface.
Scalable application	High	Experimented with Dublin, Ireland and Bologna, Italy. Can be scaled to any large-scale city. Scalability is ensured by distributed reasoning and stream computing.
System evaluation and validation of results	High	Accuracy and scalability analysis of diagnosis and prediction have been assessed against standard technologies, all reported in [3, 4, 6, 13].
Novelty in applying semantic technology	High	Semantic diagnosis and prediction are new reasoning techniques, combining semantic web technologies, stream processing and various AI techniques. Similarly the way traffic analysis and exploration is achieved is innovative.
Functionality goes beyond information retrieval	High	STAR-CITY requires strong knowledge correlation and reasoning over heterogenous data.
Commercial potential	High	Strong push and interest from some transportation department as well as IBM products (e.g., IBM Intelligent Operations Center).
Contextual information for ratings or rankings	High	All results from STAR-CITY i.e., analysis, diagnosis, exploration and prediction are context-aware. Confidence score are evaluated for diagnosis, and ranking measures are estimated for context similarity.
Multimedia documents	-	N.A
Use of dynamic data	High	Stream data from city sensors is the main characteristics of STAR-CITY.
Accurate results (i.e. use ranking)	High	Diagnosis results are ranked by confidence. Exploration is driven by semantic matching / ranking. Prediction is filtered through consistent rules.
Support for multiple languages and accessibility on a range of devices	-	N.A

Table 3: Additional Semantic Web Challenge Criteria Meet STAR-CITY.