Towards Total Traffic Awareness

Chenjuan Guo† Christian S. Jensen‡ Bin Yang†

† Department of Computer Science, Aarhus University, Denmark

‡ Department of Computer Science, Aalborg University, Denmark

† {cguo, byang}@cs.au.dk, ‡ csj@cs.aau.dk

ABSTRACT

A combination of factors render the transportation sector a highly desirable area for data management research. The transportation sector receives substantial investments and is of high societal interest across the globe. Since there is limited room for new roads, smarter use of the existing infrastructure is of essence. The combination of the continued proliferation of sensors and mobile devices with the drive towards open data will result in rapidly increasing volumes of data becoming available. The data management community is well positioned to contribute to building a smarter transportation infrastructure. We believe that efficient management and effective analysis of big transportation data will enable us to extract transportation knowledge, which will bring significant and diverse benefits to society. We describe the data, present key challenges related to the extraction of *thorough*, *timely*, and *trustworthy* traffic knowledge to achieve total traffic awareness, and we outline services that may be enabled. It is thus our hope that the paper will inspire data management researchers to address some of the many challenges in the transportation area.

1. INTRODUCTION

Transportation adversely affects many people's daily lives, and increasingly so. For example, as cities continue to grow, congestion gets worse and affects more and more people. And emissions from vehicles are responsible for high concentrations of airborne particles that increasingly threaten the health of people. For example, air pollution is believed to have contributed to as many as 1.2 million and 600,000 premature deaths in China and India in 2010[1]. Emissions also contribute to the greenhouse effect associated with the accelerating global warming that threatens to considerably affect the conditions for life on Earth.

Rapidly growing volumes of data that captures the state of a transportation infrastructure are becoming available, due to several developments. Infrastructure users increasingly carry mobile devices capable of contributing data. The infrastructure is increasingly being instrumented with data acquisition devices, e.g., infrared counting devices, Bluetooth and Wi-Fi base stations that "see" mobile devices, and cameras. As the movement towards open public data continues, it is a safe bet that such data will become available in large volumes.

An important and challenging goal is to use this data to achieve total traffic awareness. Individual data sources generally cover only a limited part of a transportation infrastructure. To enable global awareness of an entire infrastructure, the integration, or fusion, of multiple and diverse data sources is essential. To enable up-to-date awareness, new query processing techniques are called for that are capable of ingesting rapid streams of data, reflecting the data in query results with near-zero latency. The resulting total traffic awareness enables improved as well as new applications and services.

The availability of very interested stakeholders and large volumes of data allows empirical evaluations of the feasibility, effectiveness, and efficiency of data management proposals.

We note that the transportation setting is markedly different from that of the so-called *smart dust*, which was proposed in the beginning of the 1990's and gained substantial attention in the early 2000's. A key idea was to disperse large quantities of tiny sensing and communication devices in some environment, e.g., a rain forest, in order to monitor that environment. The devices would organize into a wireless sensor network that would stream data to users. While this is a compelling vision, the sizes of deployments are tiny. We are (luckily) unaware of any rain forests having been littered by large quantities of devices. In contrast, the "sensor network" of transportation is already up and running, and the continued operation of cities increasingly depend on the effective use of the data being generated.

We proceed to characterize available transportation data in Section 2. Then we describe challenges inherent in achieving total traffic awareness in Section 3, and

[1] http://tinyurl.com/bl48fq2

Types	Techniques	Accuracy	Cost	Dynamic Properties	Objects
Individual	CANBus	High	Low	Fuel consumption	Moving, stationary
	GNSS	High	Low	Instantaneous velocities, latitude-longitude locations	
Collective	PS	Low	Low	Travel times, average velocities, traffic flow	Stationary
	Cameras	High	High		
	LDs	High	High		
Media	LBSNs	Low	Low	Instant events, scheduled events, weather conditions	Moving, stationary
	Radio	High	Low		
	Web	High	Low		

Table 1: Dynamic Data Gathering Techniques

we outline applications and services that this enables in Section 4.

2. TRANSPORTATION DATA

Transportation data describes properties of stationary and moving objects that influence travel. Stationary objects include elements of a transportation infrastructure, e.g., road segments, intersections, points of interest (POIs), and regions of interest (ROIs). Moving objects includes vehicles, pedestrians, and location-based social network (LBSN) users. Both types of objects have *static* and *dynamic* properties.

Stationary objects have static properties, which may be described in *static data* sources. For example, the length, speed limit, and toll cost of a road segment may be recorded in digital maps and web pages of road authorities. The management of such data calls for spatial-data integration, e.g., location entity matching [1] and geospatial data fusion [2]. Moving objects also have static properties, such as, sizes, weights, or engine types of vehicles.

Dynamic properties of stationary and moving objects are described by *dynamic data* that can be gathered in different ways. For example, an important dynamic property of a road segment is its time-dependent travel time distribution across a day and a week, which can be obtained from data collected by Bluetooth and Wi-Fi sensors installed along the roads. We categorize three types of dynamic data gathering techniques in Table 1.

Individual gathering techniques capture individual moving objects' dynamic properties. A controller area network bus (CANBus) is an in-vehicle network that connects sensors that measure a variety of vehicle-related data, e.g., fuel consumption, at some frequency. A Global Navigation Satellite System (GNSS), e.g., GPS or Galileo, is able to capture a moving object's instantaneous velocities and latitude-longitude locations at some frequency (up to every 0.02 seconds[2]). The accuracies of the reported properties are high. For example, fuel consumption can be recorded with error rates between -2% and +2%[3].

The data collected by individual moving objects also relates to different stationary objects. Thus, it is possible to infer travel times or fuel consumptions associated with the traversals of road segments.

Collective gathering techniques are deployed at fixed locations in a transportation network, and they excel at capturing dynamic properties of stationary objects, e.g., the traffic flow of a particular road segment. Presence sensing (PS) techniques (e.g., Bluetooth, Wi-Fi, RFID, and Infrared), cameras, and loop detectors (LDs) are able to detect *occurrences* of moving objects at fixed locations where devices are deployed. Given the distance between, e.g., two cameras, and timestamps of vehicles' occurrences, the average speeds of road segment can be derived.

These techniques vary according to the fraction of objects they are able to detect. Cameras and LDs are able to capture almost all moving objects that pass by, while PS techniques are only able to capture some 20 to 30%[4] of the objects that pass by. However, deployment and maintenance costs of PS techniques are relatively low.

Media gathering techniques rely on humans to contribute data, e.g., about an accident, a scheduled event such as a football match, or the weather. These techniques generally report on dynamic properties of stationary and moving objects.

Location-Based Social Networks (LBSNs) are good at capturing instant and scheduled events, e.g., in the form of check-ins and user-generated content (*UGGC*) such as geo-tagged tweets and photos. The density of captured events depends on the density of users and on how frequently they post on LBSNs. The accuracy can vary greatly.

Radio stations can broadcast data on instant and sched-

[2] http://tinyurl.com/m7wqerc

[3] http://tinyurl.com/op6trtx

[4] http://tinyurl.com/low8vg8

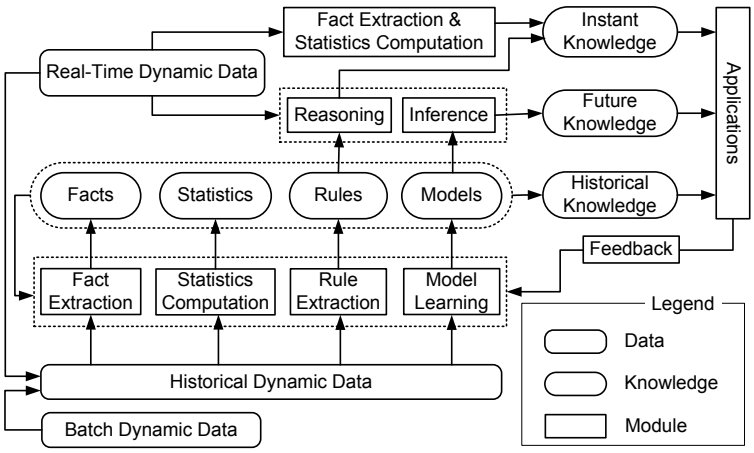

Figure 1: From Transportation Data to Thorough, Timely, and Trustworthy Transportation Knowledge

uled events and weather conditions. The accuracy is relatively high, and the deployment cost is relatively low. The web is also a good resource of schedules events and weather conditions, and accuracy is high and cost is low.

Finally, dynamic data can be classified into *real-time*, *batch*, and *historical* data. *Real-time* data is delivered immediately after being collected, *batch* data is accumulated and then delivered in batches according to some protocol, and *historical data* is delivered some time after it is collected.

3. TOTAL TRAFFIC AWARENESS

To support different applications and services, we envision a system that transforms data into transportation knowledge; see the architecture in Figure 1.

3.1 Knowledge Representation and Properties

Transportation knowledge takes the forms of *facts*, *statistics*, *rules*, and *models*. We assume that transportation data is cleaned and pre-processed before being passed through the modules in Figure 1.

Facts describe stationary and moving objects' dynamic properties, such as the average velocity of a vehicle passing through a segment. Facts are the product of integration and consolidation of dynamic data across different data sources. Thus, the *fact extraction* module integrate data elements that describe the same property of an object but are stored in different sources. It is often appropriate to associate facts with an estimated reliability.

Statistics are temporal and spatial aggregations on dynamic data and facts, such as the distribution of travel times with which vehicles pass through a road segment or the frequent routes that a user traverses between home and work during April. The *statistics computation* module must ensure the integrity of data and facts while describing them more compactly.

Rules make it possible to infer stationary-object properties from the properties of moving objects, and vice versa. For example, if at least $x\%$ of detected vehicles having velocities smaller than y then a road segment is considered as congested. And if a road segment is congested, the time needed to traverse it exceeds z minutes. Rules are produced by the *rule extraction* module.

Models describe stochastic processes that represent possible evolutions of objects across time. For example, given the current vehicle density at a road segment, a model may predict possible densities at the segment and adjacent segments 15 minutes into the future. Given an object's past trajectories, a model may predict the the object's future trajectories. The *model learning* module derives such models that are subsequently fed with real-time dynamic data to infer possible dynamic, future properties.

Next, transportation knowledge can be classified according to its temporal aspect. **Historical knowledge** describes past traffic and is obtained from the historical transportation data. **Instant knowledge** describes current or near-past traffic. It is derived by means of *fact extraction*, *statistics computation*, and *reasoning* from streaming real-time, dynamic data. Examples include the current vehicle density of a road segment and the congestion status of a road segment. **Future knowledge** infers future traffic, e.g., whether a segment is congested or clear 15 minutes from now. This knowledge is obtained from models and real-time dynamic data.

Three aspects are essential to achieve total traffic awareness. First, *thoroughness* relates to the spatial, temporal, and property coverage. Spatial coverage is thorough if the knowledge covers an entire transportation infrastructure; temporal coverage is thorough if the knowledge covers an entire period of interest; and property coverage is thorough if the knowledge fully covers the

traffic properties of interest (e.g., travel times, fuel consumption). Next, *timeliness* implies that the available knowledge is up-to-date. For example, a self-driving car requires knowledge as to whether it can go across an intersection to be at most milliseconds old, while a few seconds of delay may be acceptable for a continuous routing service. Third, *trustworthiness* means that the knowledge is sufficiently accurate and reliable for its use, even if it is derived from inaccurate and uncertain data. The knowledge should come with a quantification of its accuracy and reliability.

3.2 Challenges

To achieve total traffic awareness, it is necessary to effectively and efficiently utilize the available static and dynamic transportation data. Table 2 offers an overview of key challenges.

Thoroughness: Any single type of dynamic data is unable to offer thoroughness by itself. For example, camera data can only cover the locations where cameras are deployed. Thus, data integration is of essence. Traditional techniques, e.g., schema alignment [3] and data fusion [4], need to be adapted to the spatial-temporal aspects. Further, integration must contend with the general characteristics of transportation data. Recent techniques for *big data integration* [5,6] can be helpful. Major challenges relate to sparsity and duplicate detection.

Sparsity: Although integrating various types of dynamic data increases thoroughness, *spatial*, *temporal*, and *property* data sparsity must be addressed.

Spatial sparsity occurs because some roads lack sufficient data. For example, no data is available for a road with no PS, LDs, or camera deployment if no vehicle with a contributing GNSS devices has traversed it. Borrowing data from nearby and topologically similar roads [7, 8] may be useful for obtaining knowledge for such roads.

Extrapolation-related techniques may be used for addressing temporal sparsity. Knowledge for a road during a period when no data is available can be inferred from data from nearby periods [9], from data from nearby roads that have data for the relevant period [8], or from data from nearby or similar roads with data from nearby periods [10, 11].

Property sparsity occurs when no data captures a desired property. In such cases, it may be possible to exploit related data. For example, if only GNSS data is available for a road segment, but fuel consumption is desired, it is possible to feed the GNSS data to environmental impact models to derive fuel consumption data [12].

To fully contend with sparsity, major challenges remain. Many existing methods rely on complex mathematical optimizations that do not scale to big transportation data. Scalable solutions to solving complex optimizations are missing. Alternatively, novel problem formulations that exploit scalable techniques, e.g., scalable matrix operations, are needed.

Second, existing methods often assume static scenarios rarely consider real-time data. In contrast, our setting calls for techniques that are able to adapt to real-time data.

Third, existing techniques consider the three sparsity aspects individually, and typically rely on one type of data (primarily GNSS data). Techniques that are able to consider all the three sparsity aspects and to exploit multiple data sources in a holistic manner are called for.

Duplicate detection: If a moving object is detected by more than one data gathering technique, duplicate records are generated. For example, assume that we want to know the number of vehicles passing through a road segment during a short period. GNSS records may suggest 5, while PS records may suggest 7. Simply adding 5 and 7 is wrong if one vehicle is detected by both techniques.

Duplicate detection aims to identify the data that describes same moving objects, where the data is collected by different techniques. A simple heuristic for identifying duplicates is that if trajectories provided by different techniques are highly consistent, they may refer to the same moving object. However, the heterogeneity of trajectories is not addressed well in most of existing trajectory clustering methods [13]. Recent advances in duplicate detection in dynamic settings [14] may offer a good starting point, but cannot be applied directly.

Timeliness: Ensuring up-to-date traffic awareness presents several challenges.

Incremental maintenance: Historical knowledge, e.g., a model for predicting the travel time on a segment, is built from historical data. As such data accumulates, the historical knowledge may change. When and how the historical knowledge would change are usually unknown and cannot be predicted. This calls for an efficient and effective approach to incrementally update and maintain historical knowledge.

While we expect that it will be relatively easy to contend with facts and rules, the real challenge lies in how to maintain models. One possibility is to adopt an online learning approach, where models are updated frequently using very recent data. However, this approach may be sensitive to unscheduled traffic events, such as accidents or road construction. Another strategy may be to update the historical knowledge only when recent data disagrees significantly with the existing historical knowledge. To summarize, incremental maintenance challenges include: (1) how to efficiently and effectively identify the (dis)agreement between the existing historical knowledge and recent data; (2) how to efficiently

Properties	Challenges	Historical Knowledge	Instant Knowledge	Future Knowledge
Thoroughness	Dealing with sparsity	✓	✓	✓
	Duplicate detection	✓	✓	✓
Timeliness	Incremental maintenance	✓		
	Efficient retrieval	✓		
	Efficient processing		✓	✓
Trustworthiness	Conflict reconciliation	✓	✓	
	Veracity enhancement	✓	✓	
	Accurate prediction			✓

Table 2: Total Traffic Awareness Challenges

and effectively differentiate the two cases; and (3) how to address disagreements.

Efficient retrieval: We consider the efficient retrieval of historical knowledge. While some techniques for efficient historical knowledge retrieval do exist, we face the specific challenge that our knowledge is spatio-temporal and takes four forms.

There is a need for efficient retrieval that involves comparisons between historical knowledge and streaming real-time data. For instance, when an accident happens, it is of interest to predict the spatio-temporal extent of the congestion in the road network that is caused by the of the accident. Facts and statistics from a similar past accident that happened in a similar situation (e.g., same region, at a similar time of day, under similar weather condition) may provide reliable predictions. Rules (e.g., if an intersection is congested, its X% adjacent segments and its Y% 2-nd adjacent segments will be congested) and models (e.g., prediction of the duration of congestion) at the current accident location may also help predict the impact of the accident. How to efficiently retrieve historical facts, statistics, rules, and models that are relevant to given dynamic real-time data is an important challenge.

Efficient processing: It should be possible to generate instant and future knowledge efficiently from historical knowledge and dynamic real-time data so that up-to-date knowledge is available to applications.

The generation of facts, statistics, and traffic statuses from rules must contend with the specifics of transportation data and therefore faces challenges akin to this covered for real-time data integration [15]. As the data sources that provide transportation data keep changing, additional challenges result. Also, the solutions to the thoroughness problems should also be addressed efficiently.

Using models to predict future traffic is another challenge. Some proposals, e.g., [10, 16], address this problem. However, whether these proposals are scalable and work in a rea-time manner is unknown.

Trustworthy: Different types of data have different veracity. For example, camera and loop detectors capture all or nearly all vehicles, while Bluetooth and Wi-Fi base stations do not. Our setting is faced with large amounts of low-veracity data. However, the extracted transportation knowledge should be trustworthy. This causes a number of challenges, including the following.

Conflict reconciliation: Data from different sources may disagree on the same property of a moving object. For example, GNSS data may suggest that a vehicle travels at 50 km/h, while PS data may suggest 55 km/h. The challenge is how to derive a single, trustworthy value for a property given conflicting data, and how to quantify the trustworthiness.

Methods considering data source trustworthiness (e.g., weighted voting) [17, 18] can be adopted. Here, each type of data is associated with a weight reflecting how trustworthy it is. Using the weights, a single trustworthy value for a property can be determined. The key is how to determine the weights. Sometimes, the weight of a technique may vary due to, e.g., weather conditions. A method that can automatically assign appropriate weights and update weights when necessary is highly desired. A possible solution is to use high-veracity data, e.g., camera data, as training data, and to assign and update the weights of other types.

Veracity enhancement: Veracity enhancement aims to extract high-veracity knowledge from low-veracity data from multiple sources. An interesting solution may be to utilize social knowledge (e.g., social media data that mentions traffic, or crowdsourcing) to help increase veracity [19]. A challenge is how to utilize social knowledge in an on-line setting.

Accurate prediction: It is challenging to use historical model knowledge to accurately predict near-future traffic. This may call for novel types of models capable of accommodating traffic dynamics and that are robust enough to deal with low veracity data, e.g., by automat-

ically dropping outlier data.

4. APPLICATIONS

Total traffic awareness enables a range of applications and services, a few of which we review here.

Routing: Given a source-destination pair, a routing service suggests routes. *Eco-routing* provides routes that minimizes greenhouse gas emissions. Eco-routing needs historical knowledge to construct time-varying eco-weights that describe emissions on road segments across time and may also need instant and future knowledge to update eco-weights. Concerns for travel time and distance may also be integrated into eco-routing, yielding *multi-criteria routing* [20]. Next, *continuous-routing* utilizes current and future knowledge to provide up-to-date routes, e.g., the fastest route, from a driver's current location to the driver's destination as traffic conditions change. Finally, in *context aware, personalized routing*, different drivers are provided with different routes that best match their current preferences.

Parking: The objective is to help drivers find parking. *Capacity notification* uses current and future knowledge of parking availability to help drivers. *Nearby parking* suggests the nearest available parking. This requires historical knowledge of parking spaces that are not recorded in digital maps (e.g., on-street parking [21]) and also needs instant knowledge of current availability.

Event Response: This relates to how to respond to an event, e.g., a traffic accident or a football match. *Event detection* concerns the discovery of events. A scheduled event, e.g., a football match, can be identified from, e.g., web pages or social media. An unscheduled event, e.g., an accident, needs to be detected from instant knowledge. *Event effect* predicts the spatio-temporal effect of an event from historical, instant, and future knowledge. *Event notification* aims to notify travelers of relevant events in advance. This calls for comparison of a traveler's movement with the extent of an event. This requires instant and future knowledge.

5. REFERENCES

[1] V. Sehgal, L. Getoor, and P. Viechnicki, "Entity resolution in geospatial data integration," in *GIS*, pp. 83–90, 2006.

[2] S. Stankutė and H. Asche, "An integrative approach to geospatial data fusion," in *ICCSA*, pp. 490–504, 2009.

[3] P. A. Bernstein, J. Madhavan, and E. Rahm, "Generic schema matching, ten years later," *PVLDB*, 4(11): 695–701, 2011.

[4] X. L. Dong and F. Naumann, "Data fusion - resolving data conflicts for integration," *PVLDB*, 2(2): 1654–1655, 2009.

[5] S. Guo, X. Dong, D. Srivastava, and R. Zajac, "Record linkage with uniqueness constraints and erroneous values," *PVLDB*, 3(1): 417–428, 2010.

[6] X. Liu, X. L. Dong, B. C. Ooi, and D. Srivastava, "Online data fusion," *PVLDB*, 4(11): 932–943, 2011.

[7] T. Idé and M. Sugiyama, "Trajectory regression on road networks," in *AAAI*, 2011.

[8] B. Yang, M. Kaul, and C. S. Jensen, "Using incomplete information for complete weight annotation of road networks," *TKDE*, 2014.

[9] J. Zheng and L. M. Ni, "Time-dependent trajectory regression on road networks via multi-task learning," in *AAAI*, 2013.

[10] B. Yang, C. Guo, and C. S. Jensen, "Travel cost inference from sparse, spatio-temporally correlated time series using markov models," *PVLDB*, 6(9): 769–780, 2013.

[11] Y. Zhu, Z. Li, H. Zhu, M. Li, and Q. Zhang, "A Compressive Sensing Approach to Urban Traffic Estimation with Probe Vehicles," *IEEE Trans. Mob. Comput.*, 12(11): 2289–2302, 2013.

[12] C. Guo, Y. Ma, B. Yang, C. S. Jensen, and M. Kaul, "Ecomark: evaluating models of vehicular environmental impact," in *SIGSPATIAL/GIS*, pp. 269–278, 2012.

[13] J.-G. Lee, J. Han, and K.-Y. Whang, "Trajectory clustering: a partition-and-group framework," in *SIGMOD*, pp. 593–604, 2007.

[14] P. Li, X. L. Dong, A. Maurino, and D. Srivastava, "Linking temporal records," *PVLDB*, vol. 4, no. 11, pp. 956–967, 2011.

[15] C. Rueda and M. Gertz, "Real-time integration of geospatial raster and point data streams," in *SSDBM*, pp. 605–611, 2008.

[16] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "Driving with knowledge from the physical world," in *KDD*, pp. 316–324, 2011.

[17] A. Galland, S. Abiteboul, A. Marian, and P. Senellart, "Corroborating information from disagreeing views," in *WSDM*, pp. 131–140, 2010.

[18] X. Yin, J. Han, and P. S. Yu, "Truth discovery with multiple conflicting information providers on the web," *TKDE*, 20(6): 796–808, 2008.

[19] A. Artikis, "Heterogeneous stream processing and crowdsourcing for urban traffic management," in *EDBT*, pp. 712–723, 2014.

[20] B. Yang, C. Guo, C. S. Jensen, M. Kaul, and S. Shang, "Stochastic Skyline Route Planning Under Time-Varying Uncertainty," in *ICDE*, pp. 136–147, 2014.

[21] B. Yang, N. Fantini, and C. S. Jensen, "iPark: identifying parking spaces from trajectories," *EDBT*, pp. 705–708, 2013.