Challenges for Quality of Data in Smart Cities

PAYAM BARNAGHI and MARIA BERMUDEZ-EDO, University of Surrey
RALF TÖNJES, Hochschule Osnabrück

Categories and Subject Descriptors: H.2.8 [**Information Systems**]: Spatial-Temporal Systems; C.4 [**Computer Systems Organization**]: Dependable and Fault-Tolerant Systems and Networks; I.2.6 [**Learning**]: Knowledge Acquisition; C.3 [**Computer Systems Organization**]: Embedded and Cyber-Physical Systems

General Terms: Algorithms, Reliability, Management

Additional Key Words and Phrases: Quality of information, Internet of Things, smart cities

ACM Reference Format:
Payam Barnaghi, Maria Bermudez-Edo, and Ralf Tönjes. 2015. Challenges for quality of data in smart cities. ACM J. Data Inf. Qual. 6, 2–3, Article 6 (June 2015), 4 pages.
DOI: http://dx.doi.org/10.1145/2747881

1. INTRODUCTION

Smart cities use multimodal information coming from heterogeneous sources, including various types of the Internet of Things (IoT) data such as traffic, weather, pollution, and noise data. The smart city data usually have different quality of information (QoI). QoI of each data source mainly depends on three factors: (1) errors in measurements or precision of the data collection devices, (2) noise in the environment and quality of data communication and processing (including network-dependent quality of service parameters), and (3) granularity of the observations and measurements in both spatial and temporal dimensions. Furthermore, various environments have different requirements that will determine the efficacy of using the data in the smart city applications; some systems have energy restrictions; and some wireless networks could rely on low bandwidth or intermittent connectivity. Most smart city applications also have to deal with huge volumes of data, with high velocity, dynamicity, and a variety of types of data.

The QoI issues become more challenging when various data with different QoI are going to be integrated into an application to extract higher-level information and/or to provide actionable information to other services and applications. In some of the current smart city frameworks, the underlying information model is based on semantic

The research leading to these results has received funding from the European Commission's Seventh Framework Programme for the CityPulse project under grant agreement no. 609035.

Authors' addresses: P. Barnaghi and M. Bermudez-Edo, Institute for Communication Systems (ICS), James Clerk Maxwell Building, University of Surrey, Stag Hill Campus, Guildford, GU2 7XH, United Kingdom; emails: {p.barnaghi, m.bermudez}@surrey.ac.uk; R. Tönjes, Fakultät Ingenieurwissenschaften und Informatik, Hochschule Osnabrück, Postfach 1940, 49009 Osnabrück, Germany; email: r.toenjes@hs-osnabrueck.de.

DOI: http://dx.doi.org/10.1145/2747881

descriptions that provide an annotation model to support interoperability between different sources of information (e.g., SmartSantander [Turchi et al. 2014], Aarhus [Kolozali et al. 2014]). These models can help to represent the QoI for each data source; however, the models and annotated data are often represented as static descriptions, making them unsuitable for dynamic smart city data in which the quality can change over time.

2. CHALLENGES AND RESEARCH DIRECTIONS

QoI in smart city frameworks usually is application dependent. Depending on the requirements, the solutions for enhancing the QoI could be different. For example, if the aim of an application is to seek trends in data, large samples could leverage the QoI. In this case, techniques for aggregation, such as the symbolic aggregate approximation (SAX) algorithm [Kasetty et al. 2008], can help to create a higher-granularity representation of data by removing the high-frequency variations and allowing representation of an aggregated view of the data. However, if the interest of the application is on the latency and accuracy of the data, the increase in quality will be determined by a selection of trustable resources or a combination of data from multiple resources to provide more accurate results.

The precision of the observations and measurements can be improved by increasing the frequency and density of sampling and/or by using more accurate devices for sampling [Zhou et al. 2014]. To reduce the effects of the noisy environments, data preprocessing techniques that reduce noise can be applied [Frénay and Verleysen 2013]. To reduce the effect of the granularity of the observations and measurements, different interpolation techniques, such as linear, polynomial interpolation, and Gaussian models, can be used [Mendez et al. 2013].

To overcome the volume issues of the transmitted data in high-frequency sampling, in-network fusion techniques that only transmit outliers or dimensionality reduction techniques can be applied [Brayner et al. 2014]. The bandwidth limitation of the networks can be solved with similar techniques and also by having a hierarchical storage, where most of the information is stored in the source but less information is transmitted to the applications, using sampling or other aggregation techniques. When more granularity of the data is requested, the source could be accessed; when less granularity is requested, the sampled or aggregated data could be accessed.

Most of the data aggregation and interpolation solutions assume that the QoI at the origin of the data is higher than that at the application level. However, if information is created by combining and integrating multiple data, the accuracy of the processed information could be higher than the original data at individual sources. To keep track of the process of the information and to be able to accurately select the sources of information and the processing for each individual application, it is necessary to annotate the provenance of the information [Kolozali et al. 2014].

To describe and use the quality-related parameters of smart city data, we propose using lightweight dynamic semantics. The semantic models will provide interoperable descriptions of data and their quality and provenance attributes. The semantic annotation of quality parameters is useful for interoperability and knowledge-based information fusion. To make semantics scenario independent and to be able to fast annotate and process ontologies, we propose lightweight semantic models that contain only a few general concepts without many reasoning rules. The QoI in these models can then be updated by data processing software and APIs. As the data quality parameters of the data sources update, these changes can then be linked to and reflected in their semantic descriptions. Thus, the processing applications can access the semantic description to determine the quality parameters of the data descriptions. For the

Table I. A Summary of the Key Issues and Challenges

Issues		Challenges
Issue Type	*Issue*	
QoI	Precision	Adaptive sampling Device calibration Device accuracy
	Accuracy	Noise filtering and preprocessing provenance
	Granularity	Interpolation and spatiotemporal density
Data characteristics	Volume and velocity	Outlier detection and transmission scalability
	Variety	Interoperability and dynamic semantics
Constraints	Energy bandwidth connectivity	Resource and context-aware data collection, processing and communication hierarchical storage, and centralized versus distributed systems

aggregated and complex data that are integrated from multiple sources, the provenance parameters can help to trace the QoI parameters of each source and quality aspects of the processing algorithms and methods that are applied on the data. However, updating the dynamic semantic models and determining the quality of data at the sources, quality of the network, environment, and processing components, and monitoring their changes over different time/location dimensions is still a key challenge. A summary of challenges is shown in Table I.

Overall smart city data relies on large-scale deployment of multivendor, multiprovider devices, networks, and resources that usually operate in noisy and dynamic environments. The temporal and spatial density of sampling of data collection will have an impact on the quality of the smart city data. Different environment and network parameters add limitations such as latency and noise. Energy constraints will also have an impact on the quality of data. Semantic descriptions and annotations can be used to describe different features of the smart city data and their quality attributes. However, the conventional semantics are usually static, and their complexity hinders their application in very large scale deployments and (near) real-time applications. We propose using lightweight semantic models with provenance information and combining semantics with data interpolation and data analytic models to create dynamic semantic description of quality parameters in a smart city framework.

REFERENCES

Angelo Brayner, André L. V. Coelho, Karina Marinho, Raimir Holanda, and Wagner Castro. 2014. On query processing in wireless sensor networks using classes of quality of queries. *Information Fusion* 15, 44–55.

Benoît Frénay and Michel Verleysen. 2013. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems* 25, 5, 845–869.

Shashwati Kasetty, Candice Stafford, Gregory P. Walker, Xiaoyue Wang, and Eamonn Keogh. 2008. Real-time classification of streaming sensor data. In *Proceedings of the 20th IEEE International Conference on Tools with Artificial Intelligence*, Vol. 1. IEEE, Los Alamitos, CA, 149–156.

Sefki Kolozali, Maria Bermudez-Edo, Daniel Puschmann, Frieder Ganz, and Payam Barnaghi. 2014. A knowledge-based approach for real-time IoT data stream annotation and processing. In *Proceedings of the 2014 IEEE International Conference on Internet of Things (iThings)*. 215–222.

Diego Mendez, Miguel Labrador, and Kandethody Ramachandran. 2013. Data interpolation for participatory sensing systems. *Pervasive and Mobile Computing* 9, 1, 132–148.

Stefano Turchi, Federica Paganelli, Lorenzo Bianchi, and Dino Giuli. 2014. A lightweight linked data implementation for modeling the Web of Things. In *Proceedings of the 2014 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*. IEEE, Los Alamitos, CA, 123–128.

Xun Zhou, Shashi Shekhar, and Reem Y. Ali. 2014. Spatiotemporal change footprint pattern discovery: An inter-disciplinary survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 4, 1, 1–23.

Received December 2014; revised February 2015; accepted March 2015