

Challenges for Quality of Data in Smart Cities

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

Categories and Subject Descriptors: **[Information systems]**: Spatial-temporal systems; **[Computer systems organization]**: Dependable and fault-tolerant systems and networks; I.2.6 **[Learning]**: Knowledge Acquisition; **[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:

P. Barnaghi, M. Bermudez-Edo and R. Tönjes, 2015. Challenges for Quality of Data in Smart Cities. *ACM J. Data Inform. Quality* 0, 0, Article 0 (2015), 3 pages.
 DOI : <http://dx.doi.org/10.1145/0000000.0000000>

1. INTRODUCTION

Smart cities use multi-modal 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: a) errors in measurements or precision of the data collection devices; b) noise in the environment and quality of data communication and processing (including network dependant quality of service -QoS- parameters); and c) 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; some wireless networks could rely on low bandwidth or intermittent connectivity. Most of the smart city applications also have to deal with huge volumes of data, with high velocity, dynamicity and variety of types of data.

The QoI issues become more challenging when various data with different QoI are going to be integrated in 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 descriptions which provide an annotation model to support interoperability between different sources of information (e.g. SmartSantander [Turchi et al. 2014], Aarhus

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.

Author's addresses: P. Barnaghi and M. Bermudez-Edo, Institute for Communication Systems, (ICS), Electronic Engineering Department, University of Surrey, Guildford, GU2 7XH United Kingdom; R. Tönjes, Fakultät Ingenieurwissenschaften und Informatik, Hochschule Osnabrück, Postfach 1940 49009 Osnabrück
 Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.

© 2015 ACM 1936-1955/2015/-ART0 \$15.00
 DOI : <http://dx.doi.org/10.1145/0000000.0000000>

[Kolozali et al. 2014]). These models can help to represent the QoI for each data source; but the models and the annotated data are often represented as static descriptions which make them unsuitable for dynamic smart city data in which the quality can change over time.

2. CHALLENGES AND RESEARCH DIRECTIONS

Quality of information in smart city frameworks is usually application dependent. Depending on the requirements the solutions for enhancing the QoI could be different. For example, if the aim of an application is seeking for trends of the data, large samples could leverage the QoI. In this case, techniques for aggregation such as 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 to represent an aggregated view of the data. Whereas if the interest of the application is on latency and accuracy of the data, the increase in the quality will be determined by 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]. In order to reduce the effects of the noisy environments, data pre-processing techniques that reduce noise can be applied [Frénay and Verleysen 2013]. In order 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 and 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. In order 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].

In order to describe and use the quality related parameters of the 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. In order 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. So the processing applications can access the semantic description to determine the quality parameters of the data descriptions. For the 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.

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 spatio-temporal density
Data characteristics	volume	outlier detection and transmission
	velocity	scalability
	variety	Interoperability and dynamic semantics
Constraints	energy bandwidth connectivity	resource and context-aware data collection, processing and communication hierarchical storage and centralised vs. distributed systems

However, updating the dynamic semantic models and determining the quality of data at the sources, quality of the network and environment and also the processing components and monitoring their changes over different time/location dimensions is still a key challenge. A summary of challenges is shown in the table I.

Overall smart city data relies on large-scale deployment of multi-vendor, multi-provider 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 LV 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 (2014), 44–55.
- Benoît Fréney and Michel Verleysen. 2013. Classification in the presence of label noise: a survey. (2013).
- Shashwati Kasetty, Candice Stafford, Gregory P Walker, Xiaoyue Wang, and Eamonn Keogh. 2008. Real-time classification of streaming sensor data. In *Tools with Artificial Intelligence, 2008. ICTAI'08. 20th IEEE International Conference on*, Vol. 1. IEEE, 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 *IEEE International Conference on Internet of Things (iThings) (IEEE iThings)*.
- Diego Mendez, M Labrador, and K Ramachandran. 2013. Data interpolation for participatory sensing systems. *Pervasive and Mobile Computing* 9, 1 (2013), 132–148.
- Stefano Turchi, Federica Paganelli, Lorenzo Bianchi, and Dino Giuli. 2014. A lightweight linked data implementation for modeling the Web of Things. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on*. IEEE, 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 (2014), 1–23.