



Review

Data quality in internet of things: A state-of-the-art survey

Aimad Karkouch^{a,*}, Hajar Mousannif^b, Hassan Al Moatassime^a, Thomas Noel^c^a OSER research team, Computer Science Department, FSTG, Cadi Ayyad University, Morocco^b LISI Laboratory, Computer Science Department, FSSM, Cadi Ayyad University, Morocco^c ICube Laboratory, University of Strasbourg, France

ARTICLE INFO

Article history:

Received 31 March 2016

Received in revised form

12 July 2016

Accepted 1 August 2016

Available online 3 August 2016

Keywords:

Internet of things

Data quality

Data cleaning

Outlier detection

ABSTRACT

In the Internet of Things (IoT), data gathered from a global-scale deployment of smart-things, are the base for making intelligent decisions and providing services. If data are of poor quality, decisions are likely to be unsound. Data quality (DQ) is crucial to gain user engagement and acceptance of the IoT paradigm and services. This paper aims at enhancing DQ in IoT by providing an overview of its state-of-the-art. Data properties and their new lifecycle in IoT are surveyed. The concept of DQ is defined and a set of generic and domain-specific DQ dimensions, fit for use in assessing IoT's DQ, are selected. IoT-related factors endangering the DQ and their impact on various DQ dimensions and on the overall DQ are exhaustively analyzed. DQ problems manifestations are discussed and their symptoms identified. Data outliers, as a major DQ problem manifestation, their underlying knowledge and their impact in the context of IoT and its applications are studied. Techniques for enhancing DQ are presented with a special focus on data cleaning techniques which are reviewed and compared using an extended taxonomy to outline their characteristics and their fitness for use for IoT. Finally, open challenges and possible future research directions are discussed.

© 2016 Elsevier Ltd. All rights reserved.

Contents

1. Introduction	58
2. Data and DQ in IoT	59
2.1. A new data lifecycle	59
2.2. IoT data characteristics	59
2.3. Definition of DQ and DQ dimensions	60
2.4. DQ and DQ dimensions for IoT	60
2.5. DQ dimensions for IoT domain-specific applications	61
2.6. DQ dimensions trade-offs	61
3. Factors endangering IoT DQ and their impact	62
3.1. Factors affecting IoT DQ	62
3.2. Layered distribution of factors threatening DQ	62
3.3. Impact on IoT DQ	63
3.3.1. Impact of deployment scale, failing-dirty, vandalism and environment on DQ dimensions in the IoT	63
3.3.2. Impact of resources constraints, unreliable sensors, network, security vulnerability and privacy preservation processing on DQ dimensions in the IoT	64
3.3.3. Impact of data stream processing on DQ dimensions in the IoT	65
3.3.4. Impact of different IoT problems on ease of access, access security and interpretability DQ dimensions	66
4. Manifestation of DQ problems in IoT and their symptoms	66
4.1. DQ problems' manifestation classes	66
4.2. Symptoms of DQ dimensions difficulties associated with DQ problems classes	67

* Corresponding author.

E-mail addresses: aimad.karkouch@ced.uca.ac.ma (A. Karkouch), mousannif@uca.ma (H. Mousannif), Hassan.al.moatassime@gmail.com (H. Al Moatassime), noel@unistra.fr (T. Noel).

5.	Data outliers	67
5.1.	Definition of data outliers	68
5.2.	Types of outliers	68
5.3.	Impact of outliers in the IoT	68
5.4.	Context-awareness and data outliers	69
6.	DQ enhancement approaches	69
6.1.	Outlier detection	69
6.1.1.	Outlier detection's process description	69
6.1.2.	DQ enhanced aspects	69
6.1.3.	DQ dimensions in outlier detection	69
6.2.	Interpolation	70
6.2.1.	Definition	70
6.2.2.	DQ enhanced aspects	70
6.3.	Data integration	70
6.3.1.	Definition	70
6.3.2.	DQ enhanced aspects	70
6.4.	Data deduplication	70
6.4.1.	Definition	70
6.4.2.	DQ enhanced aspects	70
6.5.	Data cleaning	70
6.5.1.	Data cleaning benefits and phases' pipeline	70
6.5.2.	Data cleaning system's general architecture	71
6.5.3.	Comparison's taxonomy	71
6.5.4.	Data cleaning techniques comparison	72
6.5.5.	Data cleaning techniques comparison overview	77
7.	Open challenges and future research directions	77
7.1.	Challenges	77
7.2.	Future research directions	78
8.	Conclusion	79
	Acknowledgment	79
	References	79

1. Introduction

The Internet of Things (IoT) is about millions of connected, communicating and exchanging objects, scattered all over the world and generating tremendous amounts of data using their sensors every single second. IoT is a new evolution of the Internet (Evans, 2011) and has many definitions depending on the chosen viewpoint. One that relates to data reports the shifting of roles in the era of IoT. Interconnected smart things will become the major data producers and consumers instead of humans. The flow of data from the physical to the digital world will extend the awareness of computers of their surroundings, thus, gaining the ability to act on behalf of humans through ubiquitous services.

IoT has and will affect many fields in our daily life both on personal and business levels (e.g. cities, homes, health, etc.). Further, it has a significant impact on society to the extent it has become a social “symbolic capital of power” (Nataliia and Elena, 2015). A taxonomy of IoT applications is presented in Gubbi et al. (2013) which, based on the type of network availability, coverage, scale, heterogeneity, repeatability, user involvement and impact, identifies four application domains: Home and personal, enterprise, utilities and mobile. Applications based on the crossing-over of physical and cyber worlds allowed by the IoT vision (e.g. Health applications, Home energy monitoring, Smart cities, Intelligent Products, etc.) have already been created and many more are expected (Aggarwal et al., 2013; Kiritsis, 2011).

Data represent the bridge that connects cyber and physical worlds. Their importance is illustrated with the emergence of IoT semantic-oriented vision (Atzori et al., 2010) which finds its utility from the need of ways to represent and manipulate the huge amount of raw data expected to be generated from the “things”. The autonomous and continuous harvesting of data by the “things” (e.g. RFID readers, sensor nodes, etc.) easily overtakes manually entered

data. It was in 2008 when the number of connected objects has already surpassed the number of persons on the planet (Aggarwal et al., 2013). Moreover, considering the predictions in National Intelligence Council (2008), Sundmaeker et al. (2010), the number of connected objects will become even greater. In fact, as predicted in (National Intelligence Council, 2008), common things of our daily life (e.g. lamps, refrigerators, food packages, etc.) will have had embedded components allowing them to communicate and become more intelligent by the year 2025. Furthermore, technological advances have impressively sharpened the “data harvesting” capabilities of embedded sensor devices resulting in more generated data and more continuous data streams from the real world. As a result, IoT has become an important catalyzer of Big Data Analytics.

Data are a valuable asset in the IoT because they give insights about a given phenomenon, person or entity which are used by applications to provide intelligent services in a ubiquitous manner. These insights are mined from the harvested data using data mining techniques and algorithms (Tsai et al., 2014). Many works (Equille, 2007; Hand et al., 2001; Hipp et al., 2001) state the importance of data quality (DQ) for data mining processes and the impact of low DQ on the validity of the results and interpretations of such processes, leading to the conclusion that DQ and accuracy should be ensured. However, many factors characterizing the IoT including deployment scale, things' constrained resources (Branch et al., 2009) and intermittent loss of connection (Zeng et al., 2011) are endangering the quality of the produced data. Many DQ problems, measurable at the level of DQ dimensions, occur as a result of such hazardous elements. One major manifestation of these deviations in DQ are Data Outliers (Branch et al., 2009; Chandola et al., 2009; Javed and Wolf, 2012; Otey et al., 2006). However, while outliers could describe errors, they can also describe rare events (Zhang et al., 2010) which represent precious information for the applications (Knox and Ng, 1998) (e.g. “unusual” high

Download English Version:

<https://daneshyari.com/en/article/494527>

Download Persian Version:

<https://daneshyari.com/article/494527>

[Daneshyari.com](https://daneshyari.com)