Data Quality assessment and enhancement on Social and Sensor Data

Gabriel R. Caldas de Aquino, Claudio Miceli de Farias, and Luci Pirmez [*]
gabriel@labnet.nce.ufrj.br claudiofarias@nce.ufrj.br luci.pirmez@gmail.com

Universidade Federal do Rio de Janeiro - Programa de Pós-Graduação em Informática
UFRJ, Rio de Janeiro, Brasil

Abstract. Smartphones are key devices in the Internet of Things paradigm. Social networking services on the Internet can use smartphones applications as data providers. The data gathered from sensors and data harvested from social networking services can be used by different applications for providing context-aware services. However, the excellence of the data oriented services depends on the Data quality (DQ). DQ is critical for decision making mechanisms. We present the problem related to DQ when dealing with social and sensor data. Also, we present and explore a framework whose objective is to evaluate and control DQ aspects when dealing with social and sensor data.

Keywords: Data Quality · Social Network · Internet of Things.

1 Introduction

The Internet of Things (IoT) paradigm embraces several types of smart devices which are composed by sensors, actuators and other devices with networking capabilities [1]. In the context of IoT, smartphones are key devices embedded with different types of sensors. Smartphones are providers of large amounts of environmental data. In addition, social networking platforms use smartphone applications to enable the interaction of their users with the social networks anywhere and anytime. This creates a pervasive channel for users to record and share their personal activities on social platforms [11] (e.g. Twitter).

Huge data repositories are produced in this scenario. Services can use data fusion [7] techniques on sensor and social networking platforms data to perform environment actuations. Also data analytics procedures can use social and sensor data in a complementary manner for enriching the data analysis. This enables the use of the data gathered from sensors and data harvested from social media to create contextual integrated services [1].

However, the excellence of the aforementioned services depends on Data Quality (DQ) aspects [5] of the social and sensor data that is consumed. Quality is deemed as a critical requirement for decision making mechanisms, applications

[*] The authors of this research paper would like to thank Kontron for providing the private cloud server Symkloud as the basic infrastructure used in this work.

and services. Data with poor DQ aspects can lead to erroneous decisions and analysis. In this work we define DQ as *a concept that refers to how well the data corresponds to the quality necessities of data consumers* [6] [5]. An interesting fact comes from the aforementioned definition: DQ refers to the fitness of which the data is perceived by its consumer. This means that DQ would hardly be seen in the same way by different users. In fact, each data consumer requires the used data to fulfill certain criteria which he presumes essential for his own tasks at hand. DQ standardizing is the process of making the data conforms certain DQ requirements by assessing and enhancing DQ. The problem of DQ assessment is commonly addressed through DQ dimensions [4]. DQ enhancement can be done through DQ enhancement techniques [5]. In this work we present and explore a framework for social and sensor DQ standardizing. The rest of this work is divided into 3 chapters: in chapter 2 we present the Data Quality Concepts; in chapter 3 we present the Framework for Social and Sensor DQ and in chapter 4 we conclude our work.

2 Data Quality Concepts

In this chapter we briefly discuss some important Data Quality (DQ) concepts. DQ is deemed as a critical requirement for decision making mechanisms, applications and services on the IoT. There are different definitions of DQ. In this work DQ can be defined as a concept that refers to how well the data corresponds to the necessities of data processing mechanisms [6] [5]. An interesting conclusion comes from the aforementioned definition: DQ refers the fitness of which data is perceived by its consumer. This means that DQ would hardly be seen in the same way by different users. In fact, each data consumer requires the used data to fulfill certain criteria which he presumes essential for his own tasks at hand. The problem of DQ assessment is commonly addressed through data quality dimensions [4]. The ISO international standard DQ model identifies several DQ characteristics in the context of Software Engineering [8]. According to [10], the dimensions of data quality data can be categorized into four semantic aspects: (i) *intrinsic* , (ii) *accessibility*, (iii) *contextual*, and (iv) *representational.*

The (i) *intrinsic data quality semantic aspect* is related to the quality of the data in relation to itself. As examples of dimensions there are: accuracy, objectivity, believability and reputation [10]. (ii) The *accessibility data quality semantic* dimensions describe how accessible the data is for data consumers. Examples can be accessibility and access security [10]. *Contextual data quality semantic* aspect is related to how appropriate the data is for its usage. As examples of dimensions can be relevancy, value-added , timeliness, completeness and amount of data [10]. *Representational data quality semantic* aspect describes how understandable and representative of the environment the data is. Examples of dimensions are interoperability, ease of understanding, concise representation and consistent representation [10]. The DQ challenges refers to difficulties affecting any DQ dimension. Such challenges can cause data to become entirely or partially unusable, since it may not meet the requirements of data consumers.

As we discussed, DQ does not relate only to data accuracy. Instead data quality problems can surpass the accuracy dimension to other dimensions such as the aforementioned.

3 Framework for Social and Sensor Data Quality

This section presents the framework for social and sensor DQ standardizing. Standardizing DQ is making the data conforms certain DQ requirements. This framework receives data as input and transforms the input data to conform the DQ requirements of a given application or system that will receive such data as the output of the framework. The framework has two components: (i) social DQ component and (ii) sensor DQ component.

The (i) Social DQ component is responsible for standardizing social-originated data according to DQ requirements. The (ii) Sensor DQ component is responsible for standardizing sensor-originated data according to DQ requirements. Both the (i) and (ii) components present two subcomponents: a first subcomponent is responsible for DQ assessment, while the second subcomponent is responsible for DQ enhancement. DQ assessment subcomponent is responsible for the DQ evaluation according to given DQ requirements. DQ enhancement subcomponent is responsible for enhancing DQ to a given DQ requirement. The framework is illustrated at Figure 1.

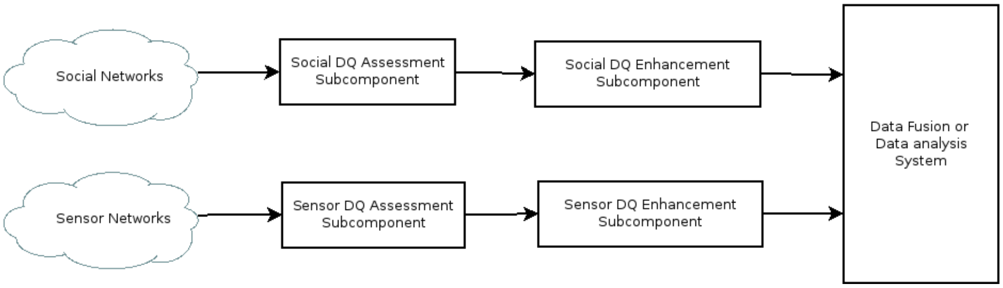
Social Networks
Social DQ Assessment Subcomponent
Social DQ Enhancement Subcomponent
Data Fusion or Data analysis System
Sensor Networks
Sensor DQ Assessment Subcomponent
Sensor DQ Enhancement Subcomponent

Fig. 1. The Framework

3.1 Social DQ Component

Social DQ component is responsible for social DQ standardizing. The first challenge for the Social DQ component is that social-originated data present significant differences according to the social media that originated the data. Differences can be if the data is structured or if it is unstructured [9], can be about the type of social interaction(the understanding of a message posted in forum thread is different from the understanding of a message posted in a Twitter discussion

[2]), etc. In other words, social data should be analyzed differently, according to the social platform that originated the message.

These differences influence the development of Social DQ analysis solutions. It is important to notice that different Social platforms may request a priori specialized DQ analysis solutions. The social DQ component is composed by two subcomponents: (i) Social DQ assessment subcomponent and (ii) Social DQ enhancement subcomponent.

The **Social DQ assessment subcomponent** aims at assessing the DQ from data originated from social networks. According to [2], Social DQ can be degraded by: (i) *Keyword ambiguity* and (ii) *Users Spamming*. (i) Keyword ambiguity is the addressing of a same context through different keywords. Even in some cases the correlation may not be obvious. Also, in the opposite, different contexts can be addressed by a common keyword. Since social data is generally collected through keyword searches, the Keyword ambiguity impacts the overall collected DQ. (ii) Users Spamming is the use of trending keywords to massively propagate a message. In some cases, such messages may not present a correlation with to the context in which the keyword is inserted. Instead, users can use trending keywords to leverage the visualization rate of the message. Spam messages can lead to the misunderstanding about the keyword context. According to [2], a key characteristic of Spam messages is its neutral tone (e.g., Check out this coupon).

The **Social DQ enhancement subcomponent** aims at processing social data to enhance the social DQ to correspond a DQ requirement of a data consumer. Given the result of the Social DQ assessment subcomponent, the Social DQ enhancement subcomponent performs the following actions: (i) accept the data to be given to the data consumer, (ii) enhance the data to the corresponding DQ requirement of the data consumer or (iii) discard the data.

3.2 Sensor DQ Component

Sensor DQ component is responsible for standardizing sensor-originated data. However, assessing and enhancing DQ for sensor data is not trivial. The sensory platforms are heterogeneous and resource-constrained. Sensor data can be originated from different kinds of sensor devices (e.g. smartphones, smart sensor networks, etc. [1]). Different sensor devices can present different data precision, data ranges, data units, hardware specifications, etc.

Regarding the sensed environment, sensory devices are placed in uncontrolled environments. In such case, misplacement, communication errors, power failure, sensor malfunction, human error or intentional misuse can potentially degrade sensor-originated DQ. The sensor DQ component is composed by two subcomponents: (i) Sensor DQ assessment subcomponent and (ii) Sensor DQ enhancement subcomponent.

The **Sensor DQ assessment subcomponent** has the objective of assessing DQ for data-collected by sensors. Much of the sensor DQ assessment can be performed on sensor data for assessing DQ according to the dimensions mentioned in the session 2. Particularly, the dimensions of believability (comparison with the correct operating bounds), completeness (missing values), free-of-error (erroneous values), consistency (over time), timeliness (delay), accuracy (deviation from true value) and precision (granularity of readings) are all important aspects of high-quality sensor data [4]. It is necessary that the sensor data well represents the events that originated the data. It implies that the data collected by multiple sensors should be processed through data fusion techniques [3] while maintaining the data consistency when representing the underlying phenomenon that originated such data.

The **Sensor DQ enhancement subcomponent** aims at the enhancement of sensor-originated DQ. Since the sensor data is constantly being integrated by data fusion procedures, it is important to perform DQ enhancement on the fly, as the data is being collected and processed. The on-the-fly data processing avoid erroneous and low quality data to propagate on fusion procedures. Also, after data fusion procedures, applying well defined DQ enhancement procedures can avoid the production of low quality data. According to [5] there are five major DQ enhancement techniques: *outlier detection*, *interpolation*, *data integration*, *data deduplication* and *data cleaning.*

(i) Outlier detection helps to improve the overall quality of datasets by making them more consistent. Moreover, outlier detection is concerned about handling instances of the unreliable datasets. Metrics used in outlier detection techniques focus on enhancing the difference between data values in order to identify outliers. (ii) Interpolation consists of inferring missing values based on other (available) values. Missing values represent gaps in available data about a certain entity or phenomena of interest for the user. As knowledge deriving processes use these datasets as input, these gaps could also lead to incomplete knowledge or wrong decisions which means that missing values could lead to a decrease in DQ. (iii) Data integration is important since social and sensor data come from different sensing platforms and different environments. In order to be used, these data need to overcome their structure differences and inconsistencies to become truly beneficial for the various services. Data integration solutions mainly focus on resolving the inconsistencies between the various data streams. (iv) Data deduplication is a data compression mechanism aiming to reduce data handling's resources consumption by reducing the amount of available data through removing of duplicate data items and replacing them with a pointer to the unique remaining copy. Data deduplication is quite simply a removal process of redundant data items. (v) Data cleaning is a process composed of 3 main phases: (i) Determination of error types, (ii) Identification of potential errors and (iii) the correction of identified (potential) errors

4 Conclusion and Future Works

In this work we presented the problem related to DQ when dealing with social and sensor data standardization. In this work we defined that standardizing DQ is making the data conforms certain DQ requirements. Also, we presented and explored a framework for social and sensor DQ standardizing. The framework we presented has the primary objective of standardizing social and sensor DQ according to an application or system DQ requirements. The proposed framework has two components: social DQ and sensor DQ components. Social DQ component is responsible for standardizing social-originated data according to DQ requirements, while Sensor DQ component is responsible for standardizing sensor-originated data according to DQ requirements. Also, each component is composed by two subcomponents: DQ assessment and DQ enhancement subcomponents. DQ assessment subcomponent is responsible for the DQ evaluation according to given DQ requirements. DQ enhancement subcomponent is responsible for enhancing DQ to a given DQ requirement.

For the realization of a framework for social and sensor DQ standardizing, a future work is to systematically study the DQ requirements for different types of applications that deal with social and sensor data. This future work aims at directing the research for solving DQ standardizing problems. Another future works that directs solutions for the problem of DQ standardizing are related to the development of techniques for assessing and enhancing sensor and social DQ aspects. We direct future works to create techniques to assess and enhance DQ considering the diverse social and sensor data inputs.

References

1. Atzori, L., Iera, A., Morabito, G.: The internet of things: A survey. Computer networks **54**(15), 2787–2805 (2010)
2. Czernek, A.: Social measurement depends on data quantity and quality - tech report (2015)
3. Farias, C., Pirmez, L., Delicato, F., Carmo, L., Li, W., Zomaya, A.Y., de Souza, J.N.: Multisensor data fusion in shared sensor and actuator networks. In: Information Fusion (FUSION), 2014 17th International Conference on. pp. 1–8. IEEE (2014)
4. Ferreira, E., Ferreira, D.: Towards altruistic data quality assessment for mobile sensing. In: Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers. pp. 464–469. ACM (2017)
5. Karkouch, A., Mousannif, H., Al Moatassime, H., Noel, T.: Data quality in internet of things: A state-of-the-art survey. Journal of Network and Computer Applications **73**, 57–81 (2016)
6. Labouseur, A.G., Matheus, C.C.: An introduction to dynamic data quality challenges. Journal of Data and Information Quality (JDIQ) **8**(2), 6 (2017)
7. Nakamura, E.F., Loureiro, A.A., Frery, A.C.: Information fusion for wireless sensor networks: Methods, models, and classifications. ACM Computing Surveys (CSUR) **39**(3), 9 (2007)
8. for Standardization/International Electrotechnical Commission, I.O., et al.: Software engineering-software product quality requirements and evaluation (square) data quality model. ISO/IEC **25012**, 1–13 (2008)
9. Tarasconi, F., Farina, M., Mazzei, A., Bosca, A.: The role of unstructured data in real-time disaster-related social media monitoring. In: 2017 IEEE International Conference on Big Data (Big Data). pp. 3769–3778 (Dec 2017). https://doi.org/10.1109/BigData.2017.8258377

10. Wang, R.Y., Strong, D.M.: Beyond accuracy: What data quality means to data consumers. Journal of management information systems **12**(4), 5–33 (1996)
11. Yerva, S.R., Jeung, H., Aberer, K.: Cloud based social and sensor data fusion. In: 2012 15th International Conference on Information Fusion. pp. 2494–2501 (July 2012)