Distributed Upstream Data Cleaning in VANET

Mohamed Ben Brahim*†, Hamid Menouar*
*Qatar Mobility Innovations Center (QMIC), Qatar University, Doha, Qatar
†HANA Research Lab, University of Manouba, Manouba, Tunisia
Email: {mohamedb, hamidm }@qmic.com

***Abstract*—The mobile road stations in Vehicular Ad hoc NETworks (VANET) are generating huge quantities of traffic-related data over wireless communication medium. For different business purposes, data has to be collected by heterogeneous distributed Road Side Units (RSUs) and pushed to traffic agencies (i.e., back-end) to be processed. Commonly, servers run a heavy processing in schema and data instance levels to ensure data transformation and data cleaning prior to storage in a production data warehouse. In particular, a comprehensive treatment is dedicated to detect and remove data redundancy and filter out or correct outlier instances. In order to alleviate these challenging tasks, a new proposal to delegate part of this data processing to the network edge in the RSUs is going to be investigated. Indeed, the RSUs are subject to potentially capture redundant data as well as defective measurement values from field data sources. Through a cooperative approach, RSUs could perform a near-real-time upstream filtering and cleaning treatment of the gathered data. Hence, an overview to tackle the upstream data cleaning and redundancy removal in a distributed fashion is presented. The proposed approach is expected to reduce the complexity of the data cleaning task and to scale better with the network size and its resulting data.**

Keywords–Distributed data cleaning; data filtering; vehicular ad hoc networks (VANET).

I. BACKGROUND

The Cyber-Physical Systems (CPSs) are empowering the establishment of future smart cities. Indeed, they cross the borders between physical and computational domains creating a better awareness of the surrounding environment beyond the visual domain. The evolution of sensing, computing, and networking technologies allowed the CPSs to ever increase the data generation causing a deluge of data that need to be managed to serve different business and social applications and services. Intelligent Transportation Systems did not deviate from the general trend and vehicles equipped with communication capabilities are able to measure and share in near-real-time traffic-related data. Similarly to other CPSs, raw data streams generated by mobile road stations might contain faulty values, redundant measurements captured by multiple road side units (RSUs), and differently structured or semi-structures data. These issues have a direct impact on the efficiency of the system as they might falsify the computing of some metrics. Hence, different techniques have been emphasized in the industrial and research communities to mitigate the negative implications of such dirty data and the resulting misleading insights. Zaho et al. [1] investigate the big data challenges resulting from urban mobility as it becomes more available. The raw urban mobility data is usually prone to inaccurate measurements due to physical or computational factors. Hence, a cleaning and pre-processing phase is usually adopted in the urban human mobitiy data mining pipeline through noise filtering, map matching, and/or road matching techniques. Shiyale and Saraf [2] investigate the dirty data cleaning in mobile wireless sensor network to enhance the overall network lifetime. The used technique is in-network and consists of enhancing the data throughput through embedded data re-transmission and leveraging spatio-temporal consistency to clean dirty data and detect outliers. Javed and Wolf [3] present a generic method using spatial and temporal characteristics to derive statistical models of continuous monitored phenomena. These models are exploited to find out defective sensor reports. NADEEF [4] is another system performing data cleaning built on top of big data processing engines like Spark [5]. Bleach [6] is a recently published system for stream data cleaning. It consists of two modules serving respectively for violation detection and violation repair. It addresses mainly real-time and high accuracy requirements. In brief, different data cleaning approaches are available in the literature. The main difference between the existing solutions is whether they address static or streaming data, they are centralized or distributed systems, and their location in the data pipeline: in-network, in the network edge, or in the central server, as depicted in Figure 1. For datasets already stored in databases, certainly techniques meant for static data, downstream, being either centralized or over cluster of workers are the relevant options. In contrast, online data streams require real-time in-network or network-edge-based methods which are usually performed in a distributed fashion.

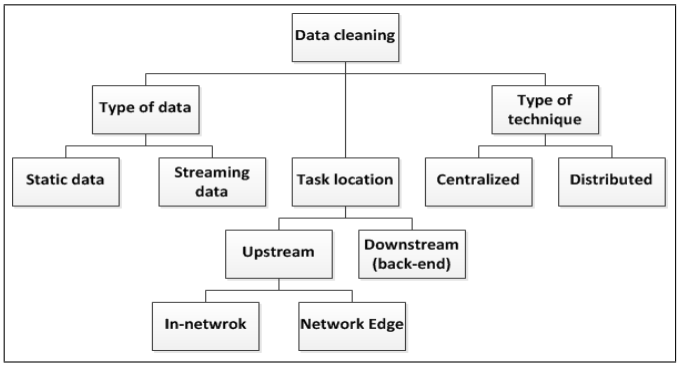

Figure 1. Taxonomy of different data cleaning techniques.

In the remainder of this paper, Section II is meant for describing data cleaning within the ETL process. Section III gives an overview of an upwind data cleaning process proposed to alleviate the data management pipeline in the back-end side. The short paper ends with a conclusion in Section IV.

II. DATA CLEANING AND ETL

The data cleaning process is required to improve data quality in different datasets. It aims at detecting and potentially fixing inconsistencies resulting from operational errors while

Copyright (c) IARIA, 2017. ISBN: 978-1-61208-572-2

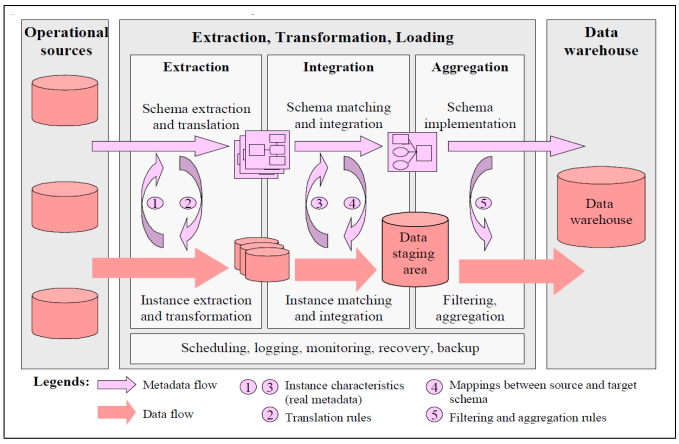

Figure 2. ETL process and data cleaning in data warehouse building [7].

populating the datasets. Data cleaning is largely studied in the well-known Extraction-Transformation-Loading (ETL) process for data warehouses. Indeed, data warehouses are usually meant for decision making and data correctness is therefore a fundamental factor. As illustrated in Figure 2, data undergo several transformation and integration tasks at instance and in schema levels. Afterwards, it is cleaned and filtered in a staging area prior to be loaded to the data warehouse [7]. Regarding the data repair, some fuzzy matching techniques are proposed to measure similarity and potentially fix erroneous data [8]. This technique of data cleaning is located at the end of the data stream for both static or live data. Hence, a considerable time delay is expected, which can reach several hours for big datasets, in order to successfully ingest the incoming data, clean it, and load it to a data warehouse.

III. Upstream Cooperative Data Cleaning

The challenges raised by the downstream data cleaning techniques are obvious. Indeed, with high velocity, high variety, and high volume of collected data, the time-to-value becomes very challenging. Hence, upwind data stream cleaning technique stands as potentially better alternative. In the context of VANET, upstream data cleaning could be performed either in-network similar to the proposed approach in wireless sensor networks [2], or could leverage the computing resources and the spatio-temporal consistency of the distributed road side units, i.e., the network edge. Indeed, as the data volume increasing rate is far exceeding the computing capacity increasing rate, it is wise to divide resources-consuming tasks into lesser complex sub-tasks and to dispatch them to multiple workers. Therefore, RSUs are considered as distributed workers which are able to cooperatively use different data mining and machine learning techniques to learn appropriate models for filtering and cleaning collected data from operational data sources in their neighborhood. As illustrated in Figure 3, mobile nodes might be in the communication range of more than single road side equipment, leading to duplicate data reception of the same packet. In addition, for multi-hop data routing techniques, the multipath scenario might occur which causes the reception of the same message by multiple collection nodes. All these scenarios, along with the inaccuracy of the measured metrics and different data structures will lead to dirty collected data. Hence, the raw collected data is not ready to serve for knowledge discovery and decision making.

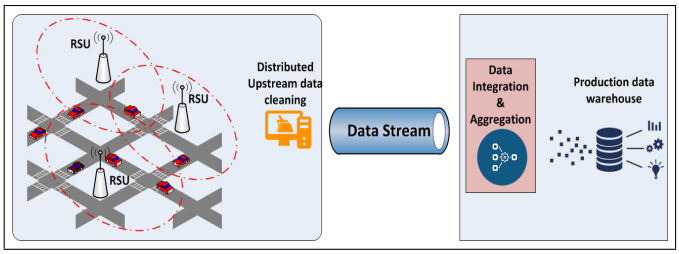

Figure 3. Distributed upstream data cleaning

An RSU could leverage some data characteristics such as spatial and temporal correlation of neighbor nodes to detect outlier values and correct them. Based on the location and time information embedded in the vehicle messages, the RSUs which are likely to receive the same message could perform a cross-check and keep a single copy of the message which solves the data redundancy problem. This could be achieved through exchange of cached messages indexes and willingness to keep a copy of each message based on available buffer size and aptitude to guarantee a data accuracy over short-time periods.

IV. Conclusion and Future Work

Data cleaning is an important task within the data management pipeline for cyber-physical systems. Since the generated data within VANET is often subject to inaccuracy and to redundant occurrences, leveraging spatio-temporal consistency and context-awareness of the distributed road side units to perform cooperative data cleaning and early data filtering is worth investigating. In the extension of this work, we will focus on defining a cooperative technique allowing the RSUs to detect and clean dirty data in the upwind of the data stream, and compare it with legacy downstream and in-network data cleaning systems.

Acknowledgment

This publication was made possible by NPRP grant #NPRP8-2459-1-482 from the Qatar National Research Fund (a member of Qatar Foundation).

References

[1] K. Zhao, S. Tarkoma, S. Liu, and H. Vo, “Urban Human Mobility Data Mining: An Overview,” IEEE International Conference on Big Data (Big Data), December 2016, pp. 1–10.

[2] K. V. Shiyale and P. D. Saraf, “Efficient Technique for Network Lifetime Enhancement by Cleaning Dirty Data,” International Journal of Science and Research (IJSR), vol. 4, no. 4, April 2015, pp. 2525–2528.

[3] N. Javed and T. Wolf, “Automated Sensor Verification using Outlier Detection in the Internet of Things,” 32nd International Conference on Distributed Computing Systems Workshops, June 2012, pp. 1–6.

[4] N. Tang, “Big Data Cleaning,” 16th Asia-Pacific Web Conference (APWeb), September 2014, pp. 13–24.

[5] “Apache Spark,” http://spark.apache.org/, 2017.06.22.

[6] Y. Tian, P. Michiardi, and M. Vukolić, “Bleach: A Distributed Stream Data Cleaning System,” arXiv:1609.05113v1 [cs.DB], September 2016.

[7] E. Rahm and H. H. Do, “Data Cleaning: Problems and Current Approaches,” Bulletin of the Technical Committee on Data Engineering, December 2000, pp. 4–13.

[8] S. Chaudhuri, K. Ganjam, V. Ganti, and R. Motwani, “Robust and Efficient Fuzzy Match for Online Data Cleaning,” Proceedings of the 2003 ACM SIGMOD international conference on Management of data, June 2003, pp. 313–324.

Copyright (c) IARIA, 2017. ISBN: 978-1-61208-572-2