Information Sciences 355–356 (2016) 90–109

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A delay-resilient and quality-aware mechanism over incomplete contextual data streams

Christos Anagnostopoulos[a], Kostas Kolomvatsos[b,*]

[a] School of Computing Science, University of Glasgow, G12 8QQ, UK
[b] Department of Computer Science, University of Thessaly, 35 100, Greece

ARTICLE INFO

Article history:
Received 20 July 2015
Revised 2 March 2016
Accepted 6 March 2016
Available online 24 March 2016

Keywords:
Incomplete multivariate context streams
Quality of streaming data
Internet of Things
Optimal stopping theory

ABSTRACT

We study the case of scheduling a Contextual Information Process (CIP) over incomplete multivariate contextual data streams coming from sensing devices in Internet of Things (IoT) environments. CIPs like data fusion, concept drift detection, and predictive analytics adopt window-based methods for processing continuous stream queries. CIPs involve the continuous evaluation of functions over contextual attributes (e.g., air pollutants measurements from environmental sensors) possibly incomplete (i.e., containing missing values) thus degrading the quality of the CIP results. We introduce a mechanism, which monitors the quality of the contextual streaming values and then optimally determines the appropriate time to activate a CIP. CIP is optimally delayed in hopes of observing in the near future higher quality of contextual values in terms of validity, freshness and presence. Our time-optimized mechanism activates a CIP when the expected quality is maximized taking also into account the induced cost of delay and an aging framework of freshness over contextual values. We propose two analytical time-based stochastic optimization models and provide extensive sensitivity analysis. We provide a comparative assessment with sliding window-centric models found in the literature and showcase the efficiency of our mechanism on improving the quality of results of a CIP.

1. Introduction

Huge volumes of sensory data in Internet of Things (IoT) environments are continuously generated as streams, which need to be analyzed on-line. Multivariate streaming data can be considered as one of the main sources of what is coined *big data*. Many IoT applications deal with multivariate contextual data coming in the form of time series like Wireless Sensors Network (WSN) data. IoT applications for, e.g., forest monitoring [34], and statistical analytics applications over large-scale data streams require efficient, accurate, and timely data analysis to facilitate (near) real-time decision-making and situational context awareness [3,4]. A contextual data stream (or context stream) contains values from contextual parameters corresponding to IoT sources, e.g., humidity sensor. IoT applications exploit all such context, for instance, to (i) infer the top-*k* recent congested segments of city road networks, or (ii) obtain regularly the highest pollution level within a time horizon in a smart city.

[*] Corresponding author. Tel.: +44 1413307252.
E-mail addresses: christos.anagnostopoulos@glasgow.ac.uk (C. Anagnostopoulos), kolomvatsos@cs.uth.gr, kostasks@di.uoa.gr (K. Kolomvatsos).

http://dx.doi.org/10.1016/j.ins.2016.03.020
0020-0255/© 2016 Elsevier Inc. All rights reserved.

1.1. Motivation & challenge

All pieces of context captured by IoT **contextual information** sources are considered as continuous context streams, where **Contextual Information Processes** (CIPs) are applied to (i) reason over incomplete data and (ii) infer new knowledge. Recent development in *big data* analytics [9] examines large amounts of contextual data to uncover hidden patterns, correlations, and other insights. With CIPs, it is possible to analyze contextual data from streams almost immediately an effort that is less efficient with more traditional business intelligence solutions. The major challenge in a stream of contextual information is that contextual data are usually imprecise, incomplete, and noisy including missing and out-of-order data. Such incompleteness is due to various errors, e.g., data interference and limitations of sensor equipment, limited WSN resources, and harsh deployment environments. The values of contextual parameters are missing, or not available, or stale. In such cases, we observe values for only a subset of contextual parameters. Hence, a CIP, which ranges from: a **data aggregation function** to an **information fusion engine**, towards to a **context inference process**, cannot be accurately evaluated. This degrades the quality of the CIP result in terms of prediction accuracy and consistent inference.

Accurate CIP results rely on the **information quality** of context stream. Stream quality is expressed by meta-information, e.g., value validity, expiration thresholds, and missingness indicators. Inaccurate observations due to missing values can be either corrected (*data imputation*) [2] or removed. However, this yields bias in the extracted knowledge and the CIP results [12]. The baseline solution is invoking a **Missing Values Substitution (MVS)** process, e.g., [6], over context streams at every time *before* the invocation of a CIP. Evidently, this imposes significant computational effort. One has to decide whether such MVS methods should be continuously invoked and at which rate. The trade-off between information quality and computational resource utilization is studied in this paper, which motivated us to introduce an intelligent mechanism for scheduling CIP invocations over incomplete context streams. The motivation here is to compensate the degree of information quality that an IoT application requires with the available computational resources, especially, when dealing with remote sensing devices. An optimally scheduled CIP over incomplete contextual information streams in order to avoid continuous calls of MVS methods establishes a mechanism that achieves the information quality levels of the application requirements. A time-optimized CIP scheduling algorithm is deemed appropriate to cope with that trade-off.

Due to incomplete data, it is difficult to determine/predict a *time instance* at which the entire set of contextual values of all context streams are present (not missing) to apply a CIP. The major research challenge here is to decide *when* to apply a CIP over context streams that are of 'good' quality. A CIP process could not be performed continuously but once within a finite time interval, which guarantees at some point quality data, thus saving computational resources. That is a CIP could be executed only when the 'necessary' information for guaranteeing quality results is available. With the term 'necessary' information we denote a degree of completeness of the context streams in order to maximize the quality of CIP results. Ideally, the maximum quality of CIP results is obtained when all values are complete/present/timely/available. The following question arises: *Given incomplete context streams, when one should activate a CIP for maximizing the quality of results by avoiding continuous call of MVS methods to save computational resources?*

2. Literature review & contribution

2.1. Literature review

We report on the CIPs that are applied over incomplete context streams and are activated once necessary information for guaranteeing quality results is available. We distinguish two basic types of a CIP over context streams: (i) CIP for Data Management and (ii) CIP for Knowledge Discovery. CIP for Data Management refers to handling, querying, scheduling, and storage of context sensory data streams. This type of CIP refers basically to data reduction and (statistical) summaries. In both cases, queries (e.g., aggregation queries and top-k queries) over context streams are executed over a summary, which refers to a compact data-structure that captures the underlying distribution of the data streams. Moreover, the well-known 'sliding window' CIP is considered as a fundamental technique for producing approximate answers to a data stream query like aggregation operators `SUM`, `AVG`, and `COUNT`. The idea behind the sliding window is to perform detailed analysis and data processing over the most recent data items. This idea has been adopted in many data stream mining and management systems [1,31]. It is worth noting that all sliding window methods invoke a CIP operator continuously over a fixed-size window. Once a piece of contextual value is missing/incomplete, then all these methods attempt to predict the missing value and then apply any CIP operator over the window.

CIP for Knowledge Discovery studies methods and algorithms for extracting knowledge from volatile context streams [7,13]. Among such types of CIPs, the on-line learning and model adaptation, concept drift detection and outliers identification have become important research topics. Pioneer contextual data stream mining processes include stream clustering [8,17], outliers detection [33], classification and prediction [24], frequent pattern [21], time series [22] and change detection [18,26]; the list is not exhaustive. Moreover, contextual information fusion processing has gained significant importance. The objective of this type of CIP is to infer the relevant states and events of the system that is being observed or activity being performed. Finally, contextual inference methods are generally applied in situational context inference [3], where inference is taken based on perceived situational knowledge.