Big Data Research 4 (2016) 1–12

Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

Big Sensor Data Applications in Urban Environments

Li-Minn Ang, Kah Phooi Seng *

School of Computing & Mathematics, Charles Sturt University, NSW 2678, Australia

ARTICLE INFO

Article history:
Received 17 July 2015
Received in revised form 13 November 2015
Accepted 24 December 2015
Available online 2 March 2016

Keywords:
Big data
Sensor-based systems
Survey
Application
Challenges

ABSTRACT

The emergence of new technologies such as Internet/Web/Network-of-Things and large scale wireless sensor systems enables the collection of data from an increasing volume and variety of networked sensors for analysis. In this review article, we summarize the latest developments of big sensor data systems (a term to conceptualize the application of the big data model towards networked sensor systems) in various representative studies for urban environments, including for air pollution monitoring, assistive living, disaster management systems, and intelligent transportation. An important focus is the inclusion of how value is extracted from the big data system. We also discuss some recent techniques for big data acquisition, cleaning, aggregation, modeling, and interpretation in large scale sensor-based systems. We conclude the paper with a discussion on future perspectives and challenges of sensor-based data systems in the big data era.

1. Introduction

Big data is a recent phenomenon with the potential to transform and enhance the values of products and services in industry and business. It is the main driver for the second economy (a concept proposed by economist W.B. Arthur which refers to the economic activities running on processors, connectors, sensors, and executors) [1]. It is estimated that by 2030, the size of the second economy will approach that of the current traditional physical economy. A definition for big data is given by [2] as "Big data is high-volume, high-velocity, and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making". An extended definition is that big data systems would involve the five *V*'s: (1) big volume of data (e.g. involving datasets of terabytes), (2) variety of data types, (3) high velocity of data generation and updating, (4) veracity (uncertainty and noise) of acquired data, and (5) big value [3]. The first four *V*'s are concerned about data collection, preprocessing, transmission, and storage. The final *V* focuses on extracting value from the data using statistical and analytical methods (e.g. machine learning algorithms, complex network theory). Big data techniques are targeted towards solving system-level problems that cannot be solved by conventional methods and technologies.

Fig. 1 shows a big data analysis pipeline [4]. The first step involves data acquisition and selecting the data required to solve the problem. For big sensor data systems (a term to conceptualize the application of the big data model towards networked sensor systems), this involves identifying and generating the required data from (multiple) sensor farms and other sources (e.g. public databases, data from social media, historical records). The second step is to perform preprocessing to obtain clean and meaningful data. This is particularly important for sensor-acquired data which is often noisy, and to remove uncertainties from the sensor data. The third step is to perform data integration, aggregation, and representation. For wireless sensor networks, the aggregation step helps in two ways. First, the volume of data is reduced for processing. Second, the process of aggregation also reduces the transmission requirements and increases the energy efficiency of battery-powered sensor nodes. The fourth step discovers new insights or knowledge from the processed data through statistical and analytical methods. The fifth step presents the data in the form of graphs or charts for human interpretation and to guide decision-making.

The number of sensors/devices available for integration into networked systems is increasing rapidly. Other than traditional sensors to measure physical quantities (e.g. temperature, pressure, light), new devices like smartphones contain embedded sensors such as microphones, cameras, accelerometers, gyroscopes, and GPS which can be used to sense a variety of data from the environment. Microphones and cameras can be used to acquire signal and image data whereas accelerometers, gyroscopes, and GPS can be used in combination to give location-based data. The internal

* Corresponding author.
E-mail addresses: kseng@csu.edu.au, jasmine.seng@gmail.com (K.P. Seng).

http://dx.doi.org/10.1016/j.bdr.2015.12.003
2214-5796/

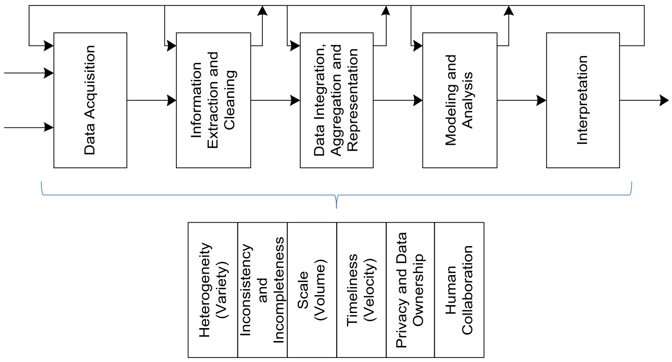

Fig. 1. The big data analysis pipeline showing major steps (top half of the figure) and characteristics that make the steps challenging (bottom half of the figure).

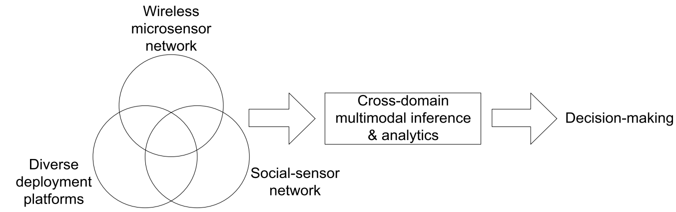

Fig. 2. The evolution of the big sensor data framework showing three inter-related branches and the required cross-domain multimodal inference and analytics for decision-making.

microprocessor clock can be used to give a timestamp on when the data was acquired. In this paper, we take a broad view of the meaning of a sensor or sensing device to describe a big sensor data system. It could be a traditional physical sensor, wearable medical sensor, smartphone, or an abstraction (e.g. energy consumption for a building, length of road network). From the data processing viewpoint, each sensor data reading contains three pieces of information which can be exploited for use in big sensor data systems: (1) measurement value, (2) timestamp, and (3) location data. The timestamp gives information on when the measurement was taken whereas the location data gives information on where in the sensing field the measurement was taken. Each sensor reading $s(x, y, t)$ can then be placed in a three-dimensional space (two dimensions of the spatial sensing field and one temporal dimension). A key characteristic of sensor-based data compared to other types of data is that it is correlated in both the spatial and temporal (spatio-temporal) domains. In the spatial domain, the sensor data forms an image snapshot of the sensing field at that particular time. In the temporal domain, each sensor produces a time series at that particular location (or nearby location in the case of mobile sensors). In a general sense, each sensor reading can be a feature vector containing several items or parameters of measurement.

We can distinguish several challenges for big sensor-based systems depending on the big data characteristics of the sensor farm deployment in terms of volume, variety, velocity, and veracity. A dense sensor farm deployment with a high sample rate would produce a "volume" challenge. The primary goal here is to ensure that there is sufficient processing power and storage available to handle the large amount of data which will be generated. Useful technologies to resolve this challenge is to employ distributed processing and storage techniques (e.g. using Hadoop, MapReduce) or cloud computing technologies. On the other hand, a sparse sensor farm deployment with a low sample rate would produce a "variety" challenge. Due to the sparseness, there would be many regions within the sensing field where there are no data readings. The primary goal here is to infer the values of the missing data points from the sensor points which are available, in combination with a variety of other correlated data sources. A complication is due to the fact that sensing devices have different sampling rates (e.g. a medical EEG sensor has very high temporal resolution in milliseconds whereas a GPS sensor has a much lower resolution in minutes). A velocity challenge would be produced for sensor network systems with real-time and latency constraints. This is often the case for event-based sensor networks. For example, a sensor network for detecting forest fires need to convey the sensed event to the base station to reach the decision maker as quickly as possible. In terms of veracity, each sensor reading comes with uncertainties not only for the measurement value. There are also uncertainties for the timestamp due to difficulties for synchronization amongst the sensors. There are also uncertainties for the location data due to difficulties for localization.

Fig. 2 shows the evolution of the big sensor data framework from three inter-related branches: wireless microsensor networks, diverse deployment platforms, and social-sensor networks. The earliest branch is the development of wireless microsensor networks, or commonly known as wireless sensor networks (WSNs) in the early 1990s. These WSN research works were initiated by DARPA which included the Distributed Sensor Networks (DSNs) and SensIT projects [65]. These early works gave sensor networks its defining capabilities like ad hoc networking, dynamic querying, reprogrammability, and multi-tasking. The early WSNs had two main characteristics. First, in terms of deployment, they were mostly confined to terrestrial or ground-based networks. Second, in terms of behavior, the sensors functioned without reference or reliance on human interaction. The next branch extended WSNs from terrestrial networks to be deployed on diverse platforms. These platforms included deployments on different mediums like underwater (underwater sensor networks), underground (underground sensor networks), aerial (satellite sensor networks) and to use different sensing modalities like speech (audio-based sensor networks), video (multimedia sensor networks), and biological signals (body sensor networks). The third evolutionary branch was due to the development of human-based social networks and smart mobile sensing devices where the human element and interaction became important. An example of a social-sensor network which will be discussed in Section 2 is the work by [31] where the Twitter social media platform was used as a distributed sensor system to serve as an early warning system for earthquake detection. The convergence of these three branches necessitates the development of a different set of big data techniques for inference and analytics. While the traditional big data problem focuses on the five V's, the big sensor data problem also requires emphasis on cross-domain and multimodal techniques to be applied towards the increasing volume and variety of networked sensors for analysis and decision-making. Cross-domain techniques refer to approaches where data can be inferred from one domain and applied in another domain. An example of this approach which will be discussed in Section 2 is the work by [16] for inferring air quality pollution where data from different domains are utilized to solve the big data problem. Multimodal techniques are required for fusion of the data from different sources (e.g. audio, speech, video, biological signals) for joint decision-making. Cross-domain and multimodal techniques will be further discussed in Section 3.

The remainder of the paper is organized as follows. Section 2 discusses several representative studies of big sensor data research in urban environments, including for air pollution monitoring, assistive living, disaster management, and intelligent transportation. Recent techniques for the big data pipeline are briefly discussed in Section 3. A discussion of future perspectives and challenges of sensor-based data systems in the big data era is given in Section 4. Section 5 concludes the paper.