Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

Big Sensor Data Applications in Urban Environments

Li-Minn Ang [a], Kah Phooi Seng [b,*]

[a] School of Computing & Mathematics, Charles Sturt University, NSW 2678, Australia
[b] School of Engineering, Edith Cowan University, WA 6027, Australia

A R T I C L E I N F O

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

Keywords:
Big data
Sensor-based systems
Survey
Application
Challenges

A B S T R A C T

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/

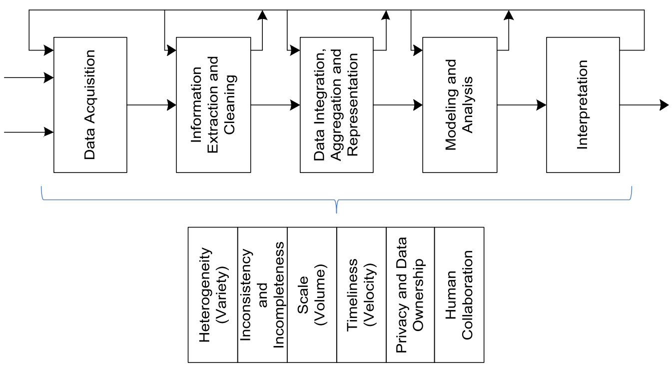

Fig. 1. The big data analysis pipeline showing major steps (top half of the figure) and characteristics that make the steps challenging (bottom half of the figure).

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.

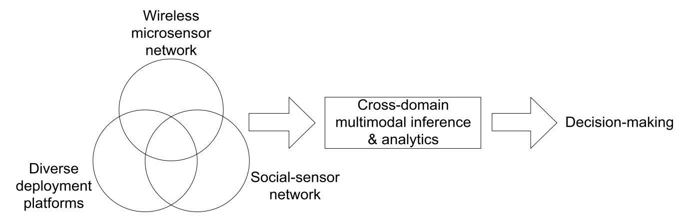

Fig. 2. The evolution of the big sensor data framework showing three inter-related branches and the required cross-domain multimodal inference and analytics for decision-making.

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.

2. Studies in big sensor data research

The current surge in big data research is driven by the needs of industries and spearheaded by companies such as Facebook, Google, LinkedIn, Twitter, and Netflix where real-time data (e.g. emails, tweets, documents, photos, videos) gathered from millions of end users (human generated sources) is used to feed large-scale analytic engines to produce additional value services such as recommender systems [5], customer analytics [6], social network analytics [7], and fraud detection [8]. It is envisaged that the next generation of big data systems will be increasingly focused on the collection, transmission, storage, and processing (analytics) of machine-generated sensor based data from sources such as networked sensor systems (e.g. Internet/Web/Network-of-Things, large scale wireless sensor systems). Much less research has been conducted in this direction and seeing how the big data paradigm for machine-generated data such as from sensor-based systems can contribute towards increasing value, although the first contributions have been made. This is becoming an increasingly important area because the volume of data from machine-generated sources is widely expected to surpass the volume of data from human-generated sources in the near future. This section surveys advancements made in the development and applications of big sensor data systems. An important focus which is usually not discussed in other surveys on networked sensors and wireless sensor systems (e.g. [9,10]) is the inclusion of the fifth *V* (i.e. how value is extracted from the big data system using the appropriate analytical and machine learning methods). The lessons learnt from the case studies and the important factors/challenges for consideration for designing and building big sensor data systems are discussed in Section 4.

2.1. Big sensor data systems for air pollution monitoring

A recent success in big sensor data systems is for inferring air quality in urban areas (e.g. cities). Air pollution is a common problem in many cities because poor air quality poses a risk to human health, particularly to people suffering from cardiovascular illnesses and to young children whose lungs are still developing. The aim is to be able to provide real-time and fine-grained air quality information (AQI index levels) to inform people and guide their daily decision-making. In urban areas, this problem is very challenging because of multiple complex factors which affect the air quality such as meteorology, traffic volume, land use, and urban structures [16]. Researchers have proposed using wireless sensor networks (WSNs) equipped with gas sensors to monitor the pollutant concentrations of CO (carbon monoxide), NO_2 (nitrogen dioxide), and O_3 (ozone) [11,12]. These gas sensors are relatively inexpensive and can be deployed on a large scale using current WSN technology. Furthermore, the sampling rate need not be very high (e.g. hourly) because the air quality would not change rapidly and there are no strict real-time constraints.

On the other hand, the problem of monitoring and detecting the aerosol pollutants such as particulate matter $PM_{2.5}$ and PM_{10} (i.e. particles with a diameter of less than 2.5 μm and 10 μm) poses more difficulty. It is important to detect these pollutants accurately because fine particulate matter is responsible for a variety of respiratory and cardiovascular diseases [13]. Unlike sensors for detecting gas pollutants, the sensors/devices for detecting aerosol pollutants are costly, not easily portable, and need a long sensing period (e.g. 1–2 hours) [16,17]. A possible solution is to apply conventional dispersion models (e.g. Gaussian Plume [14], Operational Street Canyon [15] models). However, these approaches suffer because of the difficulty to obtain the necessary modeling parameters (e.g. vehicle emission rates, street geometry, roughness coefficient of the urban surface). Thus, the only accurate way to detect and measure the air quality and aerosol pollutants content is to build a monitoring station in each area to be measured. However, these stations are costly to build (e.g. due to land cost) and maintain and it is not feasible to build multitudes of them. The authors in [16] report that Beijing only has 22 stations covering a 50 km × 50 km area. Thus, there will be many areas without monitoring stations to obtain the direct AQI information. For these regions, the big data approach is to infer the AQI index information by using the direct AQI data available from regions with monitoring stations, in combination with a variety of other indirect data sources (e.g. historical time-series data, social network tweets, real-time traffic data, city layout). Several researchers have proposed approaches using the big data model to infer the air quality for particulate matter in various cities like Beijing [16–20], New York [21], Japan [22], and Zurich [23]. Table 1 gives a summary of the different approaches showing the target city, the big data collected/used, the statistical/analytical methods used for gaining the insight/knowledge into solving the problem, and the value obtained from the big data systems.

The works in [16] and [17] describe an air quality inferring system for Beijing. In this work, the researchers divided the city into grids of cells of 1 km × 1 km. The aim is to label all cells with the AQI index level. The researchers used six AQI levels which are Good (G), Moderate (M), Unhealthy for sensitive groups (U-S), Unhealthy (U), Very unhealthy (VU), and Hazardous (H). The air quality in a grid cell is assumed to be the same throughout the cell. A grid cell which contains a monitoring station is labeled with the direct AQI level reported from the station. Five categories of indirect data features (meteorological features, traffic features, human mobility features, point-of-interest (POI) features, road network features) were extracted from the corresponding data in the cell and its eight surrounding cells. A co-training based semi-supervised learning approach was employed where unlabeled data were used to improve the inference accuracy. Two classifiers (a spatial and a temporal classifier) were built. The spatial classifier was based on a backpropagation neural network and used the static features like the road network and POI features to model the spatial correlation of air quality amongst different cells. The temporal classifier was based on a linear-chain conditional random field (CRF) and used the dynamic features like meteorological, traffic, and human mobility features to model the temporal dependency of air quality in an individual cell. The researchers reported an accuracy of 82% for the detection of PM_{10} levels and successfully inferred the AQIs for the entire Beijing in five minutes.

A different approach for inferring the air quality in Beijing was used by the researchers in [20]. The researchers observed that current coverage of AQI monitoring is limited to large cities where physical monitoring stations are built and many regions away from cities (e.g. rural towns) are under-served. To overcome this, their work proposed to use machine learning models to estimate AQI from social media data. Their key observation is that high AQI (poor air quality) in a region causes more Weibo (Sina Weibo is the most popular social media site in China) posts from that region to discuss air pollution. They found a strong positive correlation between the word “mai” (meaning haze in Chinese) with AQI levels. Their data used postings to Sina Weibo from 108 cities and collected at most 200 posts per hour per city. They also utilized the “timestamp” and “location” (GPS coordinates) data to filter off irrelevant postings. In terms of the timestamp, the posts should be within a specific one hour long period and in terms of the location, the GPS coordinates should lie within a 10-km radius circle. The researchers proposed a model based on a Markov Random Field (MRF) for AQI estimation. Other than the social media text content correlation, the MRF model also considered the spatial correlation between cities and the temporal correlation within the same city. In that sense, this is characteristic of a big sensor data model

Please cite this article in press as: L.-M. Ang, K.P. Seng, Big Sensor Data Applications in Urban Environments, Big Data Res. (2016), http://dx.doi.org/10.1016/j.bdr.2015.12.003

Table 1
Big sensor data studies for air pollution monitoring in various cities.

City	Data collected/used	Statistical/analytical method	Value	Reference
Beijing	Real-time and historical AQI levels (G, M, U-S, U, VU, H) for cells containing a monitoring station. Other variety data sources from: • Meteorological features (temperature, humidity, barometer pressure, wind speed, weather – cloudy, foggy, rainy, sunny, snowy). • Traffic features (Expectation of speeds $E(v)$, Standard deviation of speeds $D(v)$, Distribution of speeds $P(v)$). • Human mobility features (number of people arriving (f_a), number of people departing (f_l)). • Point-of-Interest (POI) features (Distribution of POIs over categories (f_n), Portion of vacant places (f_p), Changes in the number of POIs (f_c)). • Road network features (total length of highways (f_h), total length of other road segments (f_r), number of intersections (f_s)).	Co-training based semi-supervised approach using two classifiers: • spatial classifier based on backpropagation neural network. • temporal classifier based on linear-chain conditional random field (CRF).	Achieved the inferring of PM_{10} for entire Beijing in five minutes (near real-time performance) with an accuracy of 82%.	[16,17]
Beijing	$PM_{2.5}$ pollutant data from sensor network (AirCloud).	Gaussian Process Regression model.	GP inference model performed better than baseline models based on linear and cubic spline interpolation.	[18]
Beijing	Two group of data sources: • Meteorological data (temperature, humidity, wind speed, wind direct). • Pollutant data (PM_{10}, CO, NO_2, O_3, SO_2).	Backpropagation artificial neural network trained with a greedy algorithm to find the optimal combination of features from the training set.	Achieved the classification of $PM_{2.5}$ with an accuracy of 72.80%.	[19]
Beijing	Social media postings on Sina Weibo from 108 cities. Postings had a: • Time constraint (posts should be within a specific one hour long period). • Location constraint (GPS coordinates should lie within a 10-km radius circle).	Markov Random Field model that utilizes the text content in social media and the spatial-temporal correlation amongst cities and days.	Demonstrated good prediction performance for large cities. The AQI information for small cities cannot always be predicted by their nearby big cities.	[20]
New York	Two group of data sources: • Energy consumption of heating oil data from 2012 for large buildings (heating oil #2, heating oil #4, heating oil #6) collected through New York City's Local Law 84 energy disclosure mandate. • Land use and geographic data at the tax lot level from the Primary Land Use Tax Lot Output data set from the New York City Department of City Planning.	Community network detection algorithm based on the Louvain method for modularity maximization.	Graph signal model could better quantify and rank the combined impact of a building's own heating oil consumption and the consumption of its neighbors on surrounding air quality compared with a conventional method.	[21]
Japan	Time series data of $PM_{2.5}$ for 52 cities in Japan over a two year period: • Other meteorological data sources from wind speed (WS), wind direction (WD), temperature (TEMP), illuminance (SUN), humidity (HUM), and rain (RAIN).	Deep Recurrent Neural Network (DRNN) trained using a novel auto-encoder pretraining method and takes advantage of the spatial coherence (correlations) in the sensor data.	Achieved the time series prediction of $PM_{2.5}$ 12 hours ahead from the current instant and outperformed the Japanese Government $PM_{2.5}$ VENUS prediction system.	[22]
Zurich	Over 50 million UFP measurements collected by mobile sensing nodes over a period of more than two years: • Other data sources – land use and traffic data, historical measurement.	Generalized Additive Models (GAMs) to construct land-use regression (LUR) model.	Derived high resolution spatio-temporal pollution maps to show that city dwellers could reduce their exposure to UFPs by 7%.	[23]

where each city is serving as a source node, and the timestamp and location data are also utilized in solving the problem.

The researchers in [21] proposed a big data analytics model to identify clusters or communities of buildings with large $PM_{2.5}$ and NO_x amounts of emissions or "hot spots" to understand the trends of air pollution in New York City. Their model utilized heating oil consumption data from 2012 for large buildings (the burning of heavy fuel oil produces black carbon which is a key component of $PM_{2.5}$ emissions). They considered a data set where each data element is represented by a building. For each of the N data elements, there is a corresponding geographic location which was obtained from a separate set of land use and geographic data at the tax lot level. In that sense, this also resembles a big sensor data problem where each building is abstracted as a feature vector sensor source possessing spatial correlations with other neighboring buildings in the sensing space. The temporal correlations are not utilized in this case. The authors represented the built environment as a graph signal model $G = (V, W)$ where $V = \{v_0, \ldots, v_{N-1})$ is the set of nodes and W is the weighted adjacency matrix of the graph. Each data element (building) corresponds to a node v_n in the graph model and the entry $W_{i,j}$ is the weight of a directed edge that reflects the degree of relation (spatial) of the jth building to the ith building. To get the weighted adjacency matrix, the authors used a modified Gaussian dispersion plume model to define the edges between nodes. Two preprocessing steps were performed as part of the data extraction and cleaning step in the big data modeling process. A first preprocessing was performed to remove duplicate data points and data points that were incomplete or contained missing information (e.g. energy usage, square footage, geographic information). A second preprocessing was performed to identify and remove erroneous (e.g. exorbitantly high or too low energy usage) and outlier data points (e.g. top or bottom 1% of energy usage). The analysis was performed using a complex network systems technique based on the Louvain method for community detection [24]. The Louvain method is a heuristic method based on modularity maximization where modularity is a measure of the density of links inside communities as compared to the links between communities. The authors compared their graph signal model with a conventional method of ranking buildings by their weighted heating oil consumption to determine the top emitters for each pollutant ($PM_{2.5}$ and NO_x). They reported that their graph signal model could better quantify and rank the combined impact of a building's own heating oil consumption and the consumption of its neighbors on surrounding air quality compared with the conventional method. For example, the conventional method failed to identify several buildings in Manhattan where the combination of a building's own emissions and those of its neighbors together are indicative of locations where the surrounding air quality may be poor. This is because the conventional method fails to take into account the geographic locations of surrounding buildings in the analysis

The work in [22] presents an application for predicting the $PM_{2.5}$ air quality for 52 cities in Japan. The researchers used the time series data of past values of measured $PM_{2.5}$ concentrations in Japanese cities, along with other features (e.g. wind speed and rain precipitations) to predict the concentration level of $PM_{2.5}$ several hours ahead. This is another application of the big sensor data model to utilize the spatial and temporal correlations for, in this case, predicting future (sensor) values. Given a set of r sensors, the set of the resulting r time series data is given by $S = \{s_1, \ldots, s_r\}$. The objective is to predict the future values in the time series at time $\{t+1, \ldots, t+N\}$ for $s_z \in S$ given the past time series data at time $\{t, t-1, \ldots, t-L\}$. The authors proposed using a Deep Recurrent Neural Network (DRNN) trained using a novel auto-encoder pre-training method and which takes advantage of the spatial coherence (correlations) in the sensor data using the data of nearby cities in training the network. Other meteorological data sources used were from wind speed (WS), wind direction (WD), temperature (TEMP), illuminance (SUN), humidity (HUM), and rain (RAIN). The authors showed that their DRNN model achieved the time series prediction of $PM_{2.5}$ 12 hours ahead from the current instant and outperformed the VENUS system. VENUS (for Visual Atmospheric Environment Utility System) is the Japanese Government $PM_{2.5}$ prediction system based on a combination of various weather and chemical transport calculations [25].

A recent work by [23] proposed a mobile measurement system for the city of Zurich to derive accurate ultrafine particles (UFPs) pollution maps with high spatio-temporal resolution. UFPs are particles with a diameter of less than 100 nm. Their system collected a very large scale dataset of over 50 million UFP measurements using mobile sensing nodes over more than two years (from April 2012 to April 2014). The mobile measurement system consists of ten sensor nodes installed on top of public transport vehicles, which cover a large urban area (100 m × 100 m) on a regular schedule. The mobility of the sensing system enables the data to be collected with a high spatial resolution across the large area without the need for a huge number of fixed sensors. However, this comes at a cost of a reduced temporal resolution at any covered location, making it a significant challenge to derive pollution maps with a high temporal resolution at daily or hourly time scales. The authors developed land-use regression (LUR) models to produce accurate pollution maps with high spatio-temporal resolution. Their LUR model used a set of explanatory variables (land-use and traffic characteristics data) based on Generalized Additive Models (GAMs) [26] to model pollution concentrations at locations not covered by the mobile sensor nodes. The authors evaluated the dependencies between the explanatory variables and the measurements, and exploited these spatio-temporal relationships to predict the pollution levels for all locations without measurements but with available land-use and traffic information. They improved their model (decreased root-mean-square error by 26%) by incorporating historical measurements from environmental and meteorological data. They demonstrated that their system had practical value and derived high resolution spatio-temporal pollution maps to show that city dwellers could reduce their exposure to UFPs by 7% on average by not walking along the shortest path between two locations in the city but pursuing a slightly longer healthier path, which minimizes the expected exposure to UFPs. Other works on big sensor data systems for air pollution monitoring can be found in [57,58].

2.2. Big sensor data systems for assistive living

The World Health Organisation (WHO) estimates that by 2050, the number of older people on a global scale will have increased to 2 billion, which is a three-fold increase on the figure of 600 million in 2000 [27]. Consequently, and because an increasing number of elderly people wish to live independently within their own homes, new paradigms in the delivery of health and social care are required. Another growing trend for big sensor data systems is for mobile healthcare applications with the appearance of more and more wearable sensors to measure different types of health conditions (e.g. temperature, heart rate, blood pressure, pulse oximetry, electrocardiogram). This use of sensor-based technology is increasingly being seen as a solution to support assistive living, although there are several challenges to be overcome. Table 2 gives a summary of the different approaches showing the sensing device used (custom wrist sensor, smartphone, body area/sensor network) and the value obtained from the big sensor data system.

Recently, researchers in [28] have proposed a big data solution using wearable sensors to carry out continuous monitoring of elderly people, alerting caregivers when necessary, and forwarding

Table 2
Big sensor data studies for assistive living, disaster management and intelligent transportation.

City	Data collected/used	Statistical/analytical method	Value	Reference
Assistive living				
–	Wrist device with five sensors (accelerometer, temperature, thermopile, heartbeat, SpO2).	Hidden Markov Model (HMM) and Locality Sensitive Hashing (LSH) as a mechanism to learn sensor patterns for behavior recognition.	Intelligent event detection using context information to transmit only important information for analytics to reduce data volume.	[28]
–	Smartphone with embedded sensors (compass, accelerometer, gyroscope, GPS, microphone, temperature sensor, magnetometer, proximity/light sensor).	Markov Decision Process (MDP) and reinforcement learning.	Collaborative decision making among a group of sensors in close proximity for higher accuracy incident detection.	[29]
–	Wireless body area sensor network (WBASN).	Practical signal filtering algorithms for resource-constrained sensor devices (moving average, Kalman filter).	Decreases the packet loss rate down to 20% of the value obtained when compared to using a hierarchical data gathering scheme.	[30]
Disaster management				
Louisa County	"Human as sensor" through natural language via tweets.	Visualization tools to determine earthquake wave propagation.	Using social media (Twitter) as a distributed sensor system to serve as an early warning system for large scale incidents.	[31]
Japan	Spatially referenced mobile sensor data (daily GPS records from approximately 1.6 million individuals).	Machine learning technique (inverse reinforcement learning).	Model to simulate or predict population mobility in impacted cities to inform future disaster relief and management.	[32]
Japan	Auto-GPS data (9.2 billion records from more than 1 million Auto-GPS users)	Data mining using average distance traveled as key indicator.	Near real-time model of an intelligent system for optimal crisis response and evacuation management to support responsive decision-making	[33]
Intelligent transportation				
Ningbo	GPS records from approximately 4000 taxis in Ningbo, China (total of 5,521,294 GPS records).	Deep Restricted Boltzmann Machine (RBM) and Recurrent Neural Network (RNN) architecture.	Achieved congestion prediction accuracy as high as 88% within less than six minutes in a GPU-based parallel computing environment.	[34]
California	Caltrans Performance Measurement System (PeMS) database [45] (Three months traffic flow data collected every 30 seconds from over 15,000 individual detectors in freeway systems across California)	Stacked autoencoder (SAE) model trained in a layerwise greedy fashion to learn generic traffic flow features.	Achieved prediction accuracy of more than 90% for over 90% of freeways for 60-minutes prediction problem.	[35]
Shenzhen	Six months real-world data set of 14,453 taxicabs in Shenzhen, China (total of almost four billion GPS records).	Reduced 23% of passengers fares and increased 28% of drivers profits.	Reduced 60% of the total mileage to deliver all passengers and saved 41% of passengers waiting time. Recommender system based on collaborative-based filtering.	[36]

pertinent information to a big data system for analysis. Their system includes three components: a wrist device, a mobile phone, and a big data cluster. The wrist device has five sensors (an accelerometer to measure activities of the wearer, a temperature sensor to measure ambient temperature, a thermopile to measure skin temperature, and two reflective photoplethysmography sensors to measure heartbeat and SpO2 (oxygen saturation level) in the blood. The mobile phone receives the measured data from the wrist device (sent through Bluetooth Low Energy (BLE) communication) and performs intelligent behavior recognition for instant and unobstrusive care. It recognizes/infers the various states of a user (*Sleep*, *Sit*, *Stand*, *Walk*, *Run*, *Abnormal*) and controls the voice-based human–machine interaction when an anomaly is detected to avoid false-positive detection. The third component is the big data cluster/server to perform the value-based analytics. There are four potential benefits/values from this big data system: (1) personalized quality of care for each elderly person, (2) efficient use of health professional expertise (e.g. doctors and nurses), (3) provide statistical evidence for government strategic planning, and (4) reaching rural patients without proper access to healthcare.

However, a major challenge to be overcome in this big data system is the high volume of data which will be generated for storage and processing. The generation of data begins from the sensors in the wrist device. The authors report that acceleration data is sent from the wrist device every 0.1 s, skin temperature and received signal strength index (RSSI) every 2 s, heartbeat and SpO2 data every 3 s, and ambient temperature every 10 s. Every record includes a time stamp and the information stored can include various types of data (e.g. detected events, sensor readings, geolocation, voice dialog, machine triggered call, text). For a system to support 10,000 users and a replication factor of three for data redundancy, the authors estimate a daily raw data consumption of 864 GB and the consumption for one year would reach 315 TB. Even using compression technology, the authors report that this would require 4 PB of data storage to operate the system for ten years. They estimate that 222 nodes would be required in the cluster of servers and which would be very expensive, require significant power, cooling, rack space, and network port density. To deal with the high volume of data, the authors proposed an intelligent data forwarder that is embedded in each data source with

Please cite this article in press as: L.-M. Ang, K.P. Seng, Big Sensor Data Applications in Urban Environments, Big Data Res. (2016), http://dx.doi.org/10.1016/j.bdr.2015.12.003

context-aware capability. The key idea is to reduce the data at each stage in the data collection/generation stages.

From the big data model, this could also be considered to be a velocity challenge to intelligently drop redundant data from data streams while maintaining minimal information loss. The authors propose a forwarder based on a Hidden Markov Model (HMM) and Locality Sensitive Hashing (LSH) as an efficient mechanism to learn sensor patterns for human behavior recognition. The intelligent forwarders provide the remote wearable sensors with the necessary context-awareness so that the sensors transmit only important information to the big data server for analytics when certain behaviors occur and avoid overwhelming communication and data storage. This is similar to the challenges faced by event-driven wireless sensor networks where the remote sensors are equipped with smarter intelligence to detect events and transmit the sensing data only when an event is detected. In this big sensor data system, each sensing device operates independently using its context and time series information and the spatial relationships with neighboring sensors are not considered in the decision making.

Another by [29] proposed a smart collaborative mobile system for taking care of disabled and elderly people. Their system takes advantage of the sensors embedded in smartphones to monitor the status of a person based on what is happening in the environment. This is similar to the event-driven approach proposed by [28] using the intelligent data forwarder. However, there are two key differences between this system and the work in [28]. First, this system uses the sensors embedded in a smartphone as the sensing device whereas the work in [28] uses a custom wrist sensor. Second, compared to the work in [28], this system also uses information from neighboring sensors in the decision making. The main objective is to determine if a person has suffered an incident when a group of persons are doing an activity. Thus, the decision making process is collaborative based on inputs from a group of sensors in close spatial proximity. This is performed to improve the system accuracy and to predict an alarm before it happens and also reduce the amount of false positives. The alarm policy is formulated using a Markov Decision Process (MDP) and reinforcement learning. The MDP decides when to send the alarm message combining the alarm signals from various sensors. For example, if the device detects a noise above the set threshold, it could be interpreted as a distress call in a system without collaborative decision making. However, in this system, neighboring sensors within the group (i.e. in close spatial proximity) are first queried, and if their sensors also give similar data, then the alarm will not be sent because it would be interpreted that the group is moving through a noisy area. This would reduce the number of false positives. The work in [28] uses voice-based human–machine interaction for confirmation that an incident requiring attention is detected.

A wireless body area sensor network (WBASN) could be considered to be an application specific wireless sensor network of wearable biosensors to enable remote monitoring of vital health parameters (e.g. heart rate, respiration rate, pulse oximetry, blood pressure, body temperature, glucose levels, chest sounds). The work in [30] proposed using a simple and effective handoff protocol for WBASN that enables continuous monitoring of ambulatory patients at home while they recover from noncritical conditions. A focus in this work is to work within the power limitations of wearable sensors which often employ button-cell batteries to achieve compactness and small form factor. The decreased power available leads to range limitations for wireless transmission which has to be considered in the design requirements. Optimizing power consumption is a critical factor in wireless sensor-based systems powered by batteries, and this design requirement is carried forwards to apply to big sensor data systems as well. Other than having the characteristics of the five V's for traditional big data models, big sensor data models also has an additional characteristic of an E (energy efficiency) to be fulfilled. This requirement of energy efficiency should be applied at all stages in the big data pipeline whenever there is a non-rechargeable power source to be negotiated.

This work uses a two tier hierarchical network for data gathering consisting of a first tier of wearable sensors used for vital signs collection and a second tier point-to-point link between the WBASN coordinator device and a number of fixed access points (APs). In the normal case, the role of a coordinator device is to poll individual sensor devices to collect the vital signs readings before forwarding them to an AP with which it is currently associated (i.e. a conventional hierarchical routing approach where sensors transmit to coordinator/aggregator/cluster head which collects all sensor data and forwards to AP/base station). However, upon experiencing poor signal reception (the link quality is determined through the RSSI value) at the coordinator tier (e.g. when the patient moves), the AP may instruct the sensor network coordinator to forward the vital signs data through one of the wearable sensor nodes. The authors developed signal filtering algorithms that are practical for resource constrained sensor devices such as moving average and Kalman filter techniques to obtain a smoother RSSI value sequence that can be reliably referenced. In this scenario, the wearable sensor node acts as a temporary relay if the node to AP link gives a stronger signal than the coordinator to AP link. The authors showed that their approach could decrease the packet loss rate down to 20% of the value obtained when compared to using the hierarchical data gathering scheme. Other works for assistive living can be found in [59,60].

2.3. Big sensor data systems for disaster management

Sensing devices and data for big sensor data systems can take many forms; from machine-generated sensed data captured from wireless sensor motes to human-generated data from smartphone devices. This fusion of sensor network and social network has vast potential to transform human society. In this section, we will review this combined sensor-social network for disaster management applications. Disasters may occur due to meteorological events (e.g. earthquakes, hurricanes, landslides, tsunamis) or man-made events (e.g. building collapses, chemical spills, nuclear plant accidents). Table 2 gives a summary of some different approaches for disaster management showing the sensing device used, and the value obtained from the big sensor data system.

The authors in [31] discuss the emergence of social media (e.g. Twitter) as a new form of big distributed sensor network where humans act as sensors, and the generated data in the form of tweets convey relevant information with spatial and temporal characteristics reminiscent of physical wireless sensor networks. Their observation is that social media feeds often convey geographic information, as people frequently comment on events happening at their location, or refer to locations that represent momentary social hotspots. In this sense, humans perform the role of sensing and event detection. An event catches their attention, and they capture the data in the form of a written tweet and transmit to a central repository. However, one key difference between the tweet data and traditional sensor data is that the tweet information is often masked or conveyed indirectly through natural language and may not even be intended by the human source. Their study used data from the earthquake at Mineral in Louisa County (VA) in August 2011. This was the largest earthquake to hit the Eastern part of the U.S. since 1944.

The authors analyzed the response to this earthquake in Twitter by harvesting a 1% random sample of Twitter feeds for geolocation data in the period immediately following this event. The objective was to see how well the social media data could perform as a form of geosensor network. They found that the first tweet arrived 54 seconds after the event. Using only a 1% sample of tweets

Please cite this article in press as: L.-M. Ang, K.P. Seng, Big Sensor Data Applications in Urban Environments, Big Data Res. (2016), http://dx.doi.org/10.1016/j.bdr.2015.12.003

they were able to collect approximately 100 accurate geolocated tweets within two minutes of the event, and nearly 1000 such tweets within five minutes. Another key finding from this study is that the tweets could travel faster than the physical event to distant locations. An earthquake is not an instantaneous event affecting all locations within its impact zone at the same time. The seismic waves require time to propagate away from the epicenter, and this "human as sensor" big data system could serve as an early warning system for large scale incidents.

The work in [32] discusses another big sensor data system for disaster management. The authors proposed a novel Disaster Behavior Analysis and Probabilistic Reasoning System (DBAPRS) to analyze and simulate people's evacuation behaviors during the Great East Japan Earthquake and the Fukushima nuclear accident. The data for DBAPRS is obtained from approximately 1.6 million individuals throughout Japan over a one-year period (from 1 August 2010 to 31 July 2011). The authors mined this dataset of spatially referenced mobile sensor data (daily GPS records) to discover and analyze the evacuation behaviors of people during the disasters. The DBAPRS big data architecture consists of four modules: database server and visualization, discovery and analysis, learning, and probabilistic reasoning. The database server stores and manages the GPS data for all the people being tracked. For each person, the geographic location history is a series of geographic positions including longitude, latitude, and time period. The discovery and analysis module analyzes the behaviors during the disaster, and discovers long-term or short-term population evacuations. The learning module uses the discovered evacuation behaviors (movement trajectories) to train its parameters for a machine learning technique (inverse reinforcement learning) and build a probabilistic model. The probabilistic reasoning module gives the value for this system and predicts population mobility or evacuations in various cities impacted by possible disasters throughout Japan to inform future disaster relief management strategies.

A related work on tracking large population movement for discerning behavior change during crisis situations is by the researchers in [33]. In this study, the data was mined from Auto-GPS, a service provided by a leading mobile phone operator in Japan. An Auto-GPS cell phone provides a regular stream of highly accurate location data to support services that are closely linked with the user's behavior. This study used 9.2 billion records from more than 1 million Auto-GPS users for analysis. Each record contains a unique ID, timestamp, geolocation (latitude and longitude), altitude, and error level. The error level indicates the strength of the GPS signal available to the cell phone. The authors showed the potential of using Auto-GPS data as the basis for a near real-time model of an intelligent system for optimal crisis response and evacuation management to support responsive decision-making. Other works can be found in [61,62].

2.4. Big sensor data systems for intelligent transportation

The previous section has discussed the notion of "human as sensors" in big sensor data systems. Recently, using vehicles to form vehicular ad-hoc networks (VANETs) or Internet of Vehicles (IoV) has become very popular. Other than from dedicated vehicle sensor networks, there are many other sensor types in transportation environments where real-time traffic data can be sourced (e.g. onsite roadside sensors, radars, cameras, social media). This section will review big sensor data systems using the notion of vehicles as sensing elements in a large networked system for intelligent transportation, focusing on mitigating traffic congestion, predicting traffic flow, and carpooling recommendation.

Traffic congestion results in travel delays, wasted fuel consumption, and also contributes to air pollution. A possible modeling solution is to use mathematical simulation techniques (e.g. complex network theory [37]) or visualization techniques [38]. However, complex network methods generated traffic flow dynamics may not correspond to the real-world scenario. Visualization viewing techniques can show the spatio-temporal distribution of network congestion but are unable to explain the mechanism of congestion generation and predicting future trends. The authors in [34] proposed a big sensor data system for traffic congestion evolution using deep learning techniques. Deep learning algorithms use multiple-layer architectures to extract inherent features in data at different levels to discover structure in data. This study used data obtained from approximately 4000 GPS-equipped taxis in Ningbo, China from April 13, 2014 to May 9, 2014 for a total of 5,521,294 GPS records. Each GPS record contains three pieces of data: location, timestamp, and travel speed. The GPS records (updated every two minutes) were used to determine the status of a travel link (congested or not congested). The average speed of a link was calculated, and if it was lower than a threshold value (20 km/hour), then the link was marked to be congested (i.e. set as 1). For links without GPS data, the speed was calculated by using the historical records for the link. The traffic congestions for a network with N links within T time intervals is modeled as

$$\begin{bmatrix} c_1^1 & c_1^2 & \cdots & c_1^T \\ c_2^1 & c_2^2 & \cdots & c_2^T \\ \vdots & \vdots & \cdots & \vdots \\ c_N^1 & c_N^2 & \cdots & c_N^T \end{bmatrix}$$

where c_n^t represents the traffic congestion condition on the nth link at time t (a binary value of 0 or 1). The objective is to predict the elements in each row (link). The authors utilized a deep learning neural network approach to perform the prediction where the spatio-temporal correlations amongst the links are inherently learnt in the network modeling. Their solution uses a deep Restricted Boltzmann Machine (RBM) and Recurrent Neural Network (RNN) architecture. The authors showed that their model achieved a prediction accuracy as high as 88% within less than six minutes in a GPU-based (CUDA) parallel computing environment.

A related work for traffic flow prediction using deep learning is by the researchers in [35]. Accurate and timely traffic flow information has the potential to help road users make better travel decisions and alleviate traffic congestion. The evolution of traffic flow can be considered as a temporal and spatial process. The traffic flow prediction problem can be stated as follows [35]. Let X_i^t denote the observed traffic flow quantity during the tth time interval at the ith observation location in a transportation network. Given a sequence $\{X_i^t\}$ of observed traffic flow data, $i = 1, 2, \ldots, m$, $t = 1, 2, \ldots, T$, the problem is to predict the traffic flow at time interval $(t + \Delta)$ for some prediction horizon Δ. Researchers have proposed a variety of approaches for traffic flow prediction based on time series methods (e.g. autoregressive integrated moving average (ARIMA) [39], KohonenARIMA [40]) and non-parametric methods (e.g. k-nearest neighbor (k-NN) [41], Bayesian [42], neural networks [43,44]). However, the limitations of these current methods are that they were developed with only a small amount of traffic data which may not model the traffic flow features embedded in the spatio-temporal data with much accuracy. The work in [35] used a large-scale study using a data-driven approach to improve prediction accuracy. Their big sensor data model used data obtained from the Caltrans Performance Measurement System (PeMS) database [45] which consists of three months of traffic flow data collected every 30 seconds from over 15,000 individual detectors in freeway systems across California in 2013. For each detector station, the collected data are aggregated in five minute intervals. The authors used a deep learning traffic flow prediction method by training a stacked autoencoder (SAE) model to learn generic

Please cite this article in press as: L.-M. Ang, K.P. Seng, Big Sensor Data Applications in Urban Environments, Big Data Res. (2016), http://dx.doi.org/10.1016/j.bdr.2015.12.003

traffic flow features. The SAE network was trained in a layerwise greedy fashion. The spatial and temporal correlations amongst the big sensor data are inherently considered in the SAE modeling. For the 60-minutes traffic flow prediction problem, their best architecture consisted of four hidden layers, and the number of hidden units in each hidden layer was 300. The authors showed that their deep learning model gave better performance than conventional neural network models (Backpropagation, SVM, RBF). For the 60-minutes prediction, their model achieved a prediction accuracy of more than 90% for over 90% of freeways. Similar good performances were reported for the 15-minutes, 30-minutes, and 45-minutes prediction problems.

One of the biggest successes of conventional big data systems is its application in recommender systems. In general, two techniques can be employed in recommender systems: content-based filtering and collaborative-based filtering. Content-based filtering performs recommendations based on an individual user's previous responses (i.e. temporal correlations) whereas collaborative-based filtering performs recommendations based on what other similar users have preferred (i.e. spatio-temporal correlations). The authors in [36] proposed a recommender system for carpooling taxicab services. The objective of the system (termed as CallCab) is to assist passengers to find a successful taxicab ride with carpooling. In the CallCab service, a passenger can hail an occupied taxicab on streets or wait at a taxicab stand to carpool with the existing passengers. The key idea is to schedule and group related passengers (in terms of similar destination directions) into a single taxicab trip with the minimum detour mileage, thus delivering the same number of passengers with fewer taxicabs and lower mileage.

This study used six months of a real-world dataset containing GPS data from 14,453 taxicabs in the Shenzhen Municipality, with a total of almost four billion GPS records. The large GPS dataset and contexts provide the big data system with the availability to predict the future directions of occupied taxicabs (and provide recommendation services) from historical data because the taxicab trips are highly patterned. Thus, this system uses collaborative-based filtering since the recommendations are based on previous responses from other similar users. The authors showed that their system reduced 60% of the total mileage to deliver all passengers, saved 41% of passengers waiting time, reduced 23% of passengers fares, and increased 28% of drivers profits. Other works for intelligent transportation can be found in [63,64].

2.5. Discussions on case studies

The previous sections have discussed various representative studies for big sensor data research in urban environments. This section gives a critical and brief discussion on the case studies, pointing out advantages and disadvantages for different approaches, and also to relate to the issues faced by big sensor data systems. As discussed in Section 1, the big sensor data framework evolved from three inter-related branches: wireless microsensor networks or WSNs, diverse deployment platforms, and social-sensor networks. The first branch of traditional WSNs is shown by many representative works [16–19,21–23]. The work in [21] further illustrated the abstract concept of using a building as a sensor. The work in [23] showed the advantages of using a small number of mobile sensors to cover a large spatial sensing space. The second branch is concerned with using diverse deployment platforms, and not just terrestrial-based WSNs for big data sensing. Networking of satellite sensors combined with ground-based sensing have been proposed and are becoming popular for environmental monitoring and climate prediction [66,67]. Similarly, ocean-based sensor networks have been increasingly used for monitoring aquatic environments [68] and marine shellfish monitoring [69]. The third branch of social-sensor networks is shown by the works in [20] and [31]. These works used social media postings to address different big data problems. The work in [20] used postings on Sina Weibo to predict the AQI information for cities in China, whereas the work in [31] used Twitter data for earthquake prediction. These two works demonstrated the effectiveness of the "human as sensor" approach. The works in [32] and [33] used another form of human-influenced sensor data in the form of GPS records for disaster relief and crisis management. Compared to the works in [20] and [31], the works in [32] and [33] had access to a larger amount of GPS data (millions to billions of records) compared to just thousands of social media postings.

3. Big sensor data technologies

Fig. 1 shows five stages of the big data pipeline from data acquisition, to modeling and interpretation. This section will review some recent techniques for big sensor data systems. The final part will also give some discussion on cross-domain and multimodal inference and analytical techniques to be applied towards the big sensor data challenge.

3.1. Data acquisition

The big data pipeline begins with data acquisition from sensor sources. A key challenge at this stage is to reduce the amount of data to be collected or sampled from the sensing fields. If the sources have non-rechargeable power sources to be negotiated, then another significant consideration is energy efficiency. This is often the case for small battery-powered sensor nodes, and also applies to larger sensing devices like smartphones. The energy consumption for sensing is determined by its sampling rate. Thus, new techniques like compressive sensing (CS) have been shown to be effective to reduce the node energy consumption [46,47]. CS techniques work by exploiting the sparseness found in typical sensed signals where the data can be efficiently represented in some basis (e.g. Fourier, wavelet). Although traditional compression techniques can reduce the amount of data to be transmitted by removing redundancies, they still need to sample the data at high rates, and thus incurs high energy consumption for sampling. Furthermore, the compression process itself incurs additional energy and requires more computational power from sensor nodes.

CS techniques work by trading off a simpler data acquisition requirement at the node level with a heavier computational requirement to reconstruct the CS sampled data at the base station. This asymmetric arrangement is well-suited for big sensor data systems where substantial processing power (without energy constraints) is available at the central station. The authors in [48] proposed another approach for reducing node energy consumption for data collection by scheduling nodes to sleep (i.e. turn off their radios). The challenge is that sleeping nodes cannot participate in network functions (e.g. routing), with the possibility that parts of the network become partitioned and are not reachable by any node. The authors showed that their management protocol could maintain both full connectivity and higher than 90% coverage in large-scale sensor networks.

3.2. Data information extraction and cleaning

Data captured from the physical world through sensor devices tends to be noisy, incomplete, and unreliable. Traditional data cleaning techniques for conventional big data (e.g. data warehousing) do not take into account the strong spatial and temporal correlations typically present in sensor-based data types. Information about an event in sensor data is usually reflected in multiple measurement points due to overlapping areas of coverage. Thus, the inconsistency among multiple sensor measurements serves as

Please cite this article in press as: L.-M. Ang, K.P. Seng, Big Sensor Data Applications in Urban Environments, Big Data Res. (2016), http://dx.doi.org/10.1016/j.bdr.2015.12.003

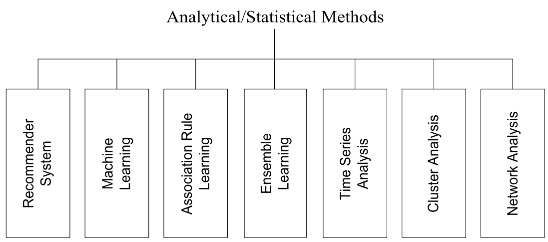

Fig. 3. Analytical/statistical methods for big data systems.

an indicator for the data quality. Some useful approaches for data cleaning for sensor data have used techniques like spatio-temporal regression and Kalman filters [49]. The authors in [50] proposed three models to detect and identify erroneous data among inconsistent observations based on the inherent structure of various sensor measurements from a group of sensors. The first model used multivariate Gaussian model to explore the correlated data changes of a group of sensors. The second model used principal component analysis (PCA) to capture the sparse geometric relationship among sensors, and the third model used kernel functions to map the original data into a high dimensional feature space prior to using the PCA model (i.e. kernel PCA). Their results demonstrated good detection rates with limited false alarms.

3.3. Data integration, aggregation and representation

The authors in [51] defined two types of sensor-based applications depending on the information that is needed at the sink: (1) functional – where only statistical summary values (e.g. maximum, average, median) are required, and (2) recoverable – where the full dataset is required. For functional applications, data integration and aggregation can be easily performed during the data collection and transmission process as part of a hierarchical data gathering tree. On the other hand, the challenges of recoverable applications pose more difficulty for data aggregation due to the lack of prior knowledge on the spatial data correlation structure. The recent work in [51] for a functional sensor application, proposed a data aggregation approach based on compressed sensing (CS). The authors showed that their CS scheme based on diffusion wavelets could achieve high fidelity recovery for aggregated sensor data while achieving significant energy savings.

3.4. Data modeling and analysis

The aim of data modeling and analysis is to derive value from the big data system and discover new insight or knowledge through analytical and statistical methods. Fig. 3 shows some analytical/statistical methods used for big data systems. In this section, we will discuss some of the techniques, and point to the reviewed case studies in Section 2 when relevant. Recommender systems are arguably the defining analytic method for big data systems. Examples of such systems are used by Amazon and Netflix to recommend books and movies to users. A recommender system contains two classes of entities (users and objects) which are grouped into two different sets (set of users $U = \{u_1, \dots, u_n\}$, and set of objects $O = \{o_1, \dots, o_m\}$). Users have preferences that must be inferred from the data. Let R be the rating (or utility) matrix, where $r(u, o)$ denotes the rating (preference) of a user u for an object o. The rating/utility matrix R is usually sparse, and the goal of a recommender system is to predict the unknown entries in the utility matrix to infer the preferences of a user. The study in [36] used a recommender system for carpooling. A recent trend for machine learning in big data is to use deep learning techniques [52]. Deep learning exploits multiple layers of information-processing in a hierarchical architecture for pattern classification and representation learning. The studies in [16,17,19,22,32,34,35] used machine learning techniques, with the studies in [34,35] using deep learning neural network models. Ensemble learning use multiple models to obtain better performance than those that could be obtained from any of the constituent models [53].

3.5. Data interpretation

To aid human interpretation and decision-making for big data systems, visualization tools can be used. Useful visualization tools include Tableau [54] for map/location-based data and Cytoscape [55] for network specific visualization.

3.6. Cross-domain and multimodal inference and analytics

An assumption made in traditional machine learning techniques is that the training and testing data are obtained from the same data domain and has the same distribution. This may not be the case in big sensor systems where data is collected and drawn from a variety of sensing sources and domains. Furthermore both real-time and historical sources may be required to be utilized. An example can be seen in the work by [16] for inferring air quality pollution where data from multiple sources and different domains are required to solve the big data problem. This necessitates the development of cross-domain machine or also known as transfer learning techniques to integrate together data from related but different information sources or coming from different historical times. The work in [70] showed an application of cross-domain learning to reduce the calibration effort of learning a model for calculating where a client device is located in a wireless network. The works in [71] and [72] proposed to apply transfer learning techniques towards solving sensor-based activity recognition in indoor environments. Even when the sensor data are collected from the same domain, they may require multimodal techniques to fuse the collective information contained in the data for decision-making. As an example, a single video stream contains two modalities (audio-based data and visual-based data) which have to be combined. Examples of multimodal fusion for multimedia analysis can be found in the survey paper by [73] which classifies the methods into three categories (rule-based methods, classification-based methods and estimation-based methods). The work in [74] further elaborated on using a linear weighted rule-based method for person identification from audio-visual sources.

4. Future perspectives and challenges

This paper has reviewed some case studies for big sensor data systems. In this section, we extract some lessons learnt from the studies, and put forward some factors/challenges for designing and building big sensor data systems.

- Spatiotemporal correlations in sensor data – the big sensor data model which treats any form of data containing [value, timestamp, location] records is useful to exploit spatiotemporal correlations in the sensing field for many forms of analytics/applications. The concept is to consider the spatial sensing field as evolving in a time series. This model can be extended for (higher) three dimensional spatial sensing applications (e.g. structural building monitoring). This will require the development of new spatiotemporal stream processing techniques.
- V's and E – Other than the 5V's, big sensor data systems also need to consider an E (energy efficiency) to be fulfilled. This requirement of energy efficiency should be applied at all stages in the big data pipeline whenever there is a non-rechargeable power source to be negotiated. This may require the development of new data gathering/routing models where

for example, sensors can itself serve as relays to transmit directly to the base station or AP if energy efficiency can be increased. The energy efficiency requirement can be considered in two ways: network energy consumption or network lifetime. The network energy consumption refers to the total energy expenditure for all sensor nodes and components (i.e. a global measure). The network lifetime is defined as the time span from the deployment to the instant when the network becomes non-functional (e.g. when *x* number of nodes die or the network becomes partitioned). In wireless sensor-based systems, optimizing the network lifetime (a local measure) is becoming the more important metric.

- Importance of "Variety" – Conventional big data systems have been mostly focused on resolving the "Volume" characteristic (e.g. through high-performance distributed processing and storage techniques), although some researchers have advocated a changing emphasis towards "Variety" [56]. From the various case studies, we see that a strong emphasis for big sensor data systems is towards using a variety of data sources or historical data to infer missing data or predict future trends in the spatial-temporal sensing field. The challenge is to find suitable models to integrate the various sources of data to solve the big data problem.
- From mathematical/simulation-driven to data-driven research – Whereas mathematical models and simulation techniques have been useful for studying the characteristics and behaviors of smaller scale systems, the move to study large-scale systems necessitate the development of new data-driven modeling techniques. Conventional mathematical and simulation models face difficulty in acquiring the correct parameters or in dealing with unpredictable/unknown factors.
- The emergence of sensor-social networks – Many studies use a combination of data from both machine-generated (e.g. GPS) and human-generated sources (e.g. Twitter). This fusion of sensor network and social network give a diversity of sources for data-driven research where new data can be inferred from one domain and applied in another domain.
- Importance of machine learning techniques – Machine learning techniques aim to develop models from data without any prior assumptions. Thus, they are a good fit for data-driven research techniques. Amongst, the machine learning approaches, the emergence of deep learning techniques compared to conventional (shallow) techniques show potential to discover hidden insights and trends in big data systems.
- The need for (near) real-time systems – Many applications (e.g. air pollution monitoring, earthquake early warning system) need (near) real-time performance to serve its function. This will drive the "Velocity" characteristic for big sensor data systems. Currently, most (if not all) research on big sensor data systems do not consider this aspect (performed offline), and research is conducted using historical or past data.

In the future, we anticipate the research and development of big sensor data systems where real-time analytics will be performed on large volumes of recently acquired data from (multiple) sensor farms, and using a number of diverse and historical sources, to produce valuable outcomes for human society. Other issues would be to do with security/privacy and the veracity challenge in big sensor data systems.

5. Conclusion

This paper has reviewed several research works for big sensor data systems in urban environments, and its applications for air pollution monitoring, assistive living, disaster management, and intelligent transportation. We have discussed how value is extracted from the big data system using analytical and statistical techniques like machine learning, recommender system, and network analysis. Many studies use techniques to exploit the spatiotemporal relationships found in big sensor data. We have also discussed how the big data pipeline can be applied towards large-scale networked sensor systems, and identified some challenges and trends for future work.

Acknowledgement

This work was partially funded by Charles Sturt University Compact Funding (A541-2006-121-25411).

References

[1] W.B. Arthur, The second economy, McKinsey Q. (Oct. 2011).
[2] http://www.gartner.com/it-glossary/big-data.
[3] T. Huang, L. Lan, X. Fang, P. An, J. Min, F. Wang, Promises and challenges of big data computing in health sciences, Big Data Res. 2 (1) (2015) 2–11.
[4] H.V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J.M. Patel, R. Ramakrishnan, C. Shahabi, Big data and its technical challenges, Commun. ACM 57 (7) (2014) 86–94.
[5] C. Sun, R. Gao, H. Xi, Big data based retail recommender system of non E-commerce, in: Proc. Int. Conf. Computing, Communication and Networking Technologies, 2014, pp. 1–7.
[6] N. Sun, J.G. Morris, J. Xu, X. Zhu, M. Xie, iCARE: a framework for big data-based banking customer analytics, IBM J. Res. Dev. 58 (5/6) (2014), 9 pp.
[7] X. Han, L. Tian, M. Yoon, M. Lee, A big data model supporting information recommendation in social networks, in: Proc. Second Int. Conf. Cloud and Green Computing, 2012, pp. 810–813.
[8] U. Srinivasan, B. Arunasalam, Leveraging big data analytics to reduce healthcare costs, IT Prof. 15 (6) (2013) 21–28.
[9] J. Yick, B. Mukherjee, D. Ghosal, Wireless sensor network survey, Comput. Netw. 52 (12) (2008) 2292–2330.
[10] L.M. Borges, F.J. Velez, A.S. Lebres, Survey on the characterization and classification of wireless sensor network applications, IEEE Commun. Surv. Tutor. 16 (4) (2014) 1860–1890.
[11] J.-H. Liu, Y.-F. Chen, T.-S. Lin, D.-W. Lai, T.-H. Wen, C.-H. Sun, J.-Y. Juang, J.-A. Jiang, Developed urban air quality monitoring system based on wireless sensor networks, in: Proc. Fifth Int. Conf. Sensing Technology, 2011, pp. 549–554.
[12] A.H. Nograles, C.P.D. Agbay, I.S.L. Flores, L. Manuel, J.B.C. Salonga, Low cost internet based wireless sensor network for air pollution monitoring using Zigbee module, in: Proc. Fourth Int. Conf. Digital Information Technology and Applications, 2014, pp. 310–314.
[13] http://www.epa.gov/airscience/air-particulatematter.htm.
[14] M. Benarie, Long-term plume models, Urban Air Pollution Modeling, 2003, pp. 258–290.
[15] S. Vardoulakis, B.E.A. Fisher, K. Pericleous, N. Gonzalez-Flesca, Modelling air quality in street canyons: a review, Atmos. Environ. 37 (2) (2003) 155–182.
[16] Y. Zheng, F. Liu, H.-P. Hsieh, U-Air: when urban air quality inference meets big data, in: Proc. 19th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2013, pp. 1436–1444.
[17] Y. Zheng, X. Chen, Q. Jin, Y. Chen, X. Qu, X. Liu, E. Chang, W-Y. Ma, Y. Rui, W. Sun, A cloud-based knowledge discovery system for monitoring fine-grained air quality, Microsoft Technical Report.
[18] Y. Cheng, X. Li, Z. Li, S. Jiang, X. Jiang, Fine-grained air quality monitoring based on Gaussian process regression, in: Lect. Notes Comput. Sci., vol. 8835, 2014, pp. 126–134.
[19] L. Xia, R. Luo, B. Zhao, Y. Wang, H. Yang, An accurate and low-cost PM2.5 estimation method based on artificial neural network, in: Proc. 20th Asia and South Pacific Design Automation Conference, 2015, pp. 190–195.
[20] S. Mei, H. Li, J. Fan, X. Zhu, C.R. Dyer, Inferring air pollution by sniffing social media, in: Proc. IEEE/ACM Int. Conf. Advances Social Networks Analysis and Mining, 2014, pp. 534–539.
[21] R.K. Jain, J.M.F. Moura, C.E. Kontokosta, Big data + big cities: graph signals of urban air pollution, IEEE Signal Process. Mag. (2014) 130–136.
[22] B.T. Ong, K. Sugiura, K. Zettsu, Dynamic pre-training of deep recurrent neural networks for predicting environmental monitoring data, in: Proc. IEEE Int. Conf. Big Data, 2014, pp. 760–765.
[23] D. Hasenfratz, O. Saukh, C. Walser, C. Hueglin, M. Fierz, T. Arn, Jan Beutel, L. Thiele, Deriving high-resolution urban air pollution maps using mobile sensor nodes, Pervasive Mob. Comput. 16 (2015) 268–285.
[24] A. Lancichinetti, S. Fortunato, Community detection algorithms: a comparative analysis, Phys. Rev. E 80 (2009).
[25] S. Sugata, T. Ohara, J. Kurokawa, M. Hayasaki, Development of air pollution forecast system (VENUS) and its validation, J. Jpn. Soc. Atmos. Environ. (2011) 49–59.

[26] T.J. Hastie, R.J. Tibshirani, Generalized Additive Models, vol. 43, Chapman & Hall/CRC, 1990.
[27] http://www.un.org/esa/population/publications/WPA2009/workingPaper.pdf.
[28] P. Jiang, J. Winkley, C. Zhao, R. Munnoch, G. Min, L.T. Yang, An intelligent information forwarder for healthcare big data systems with distributed wearable sensors, IEEE Syst. J. 99 (2014).
[29] S. Sendra, E. Granell, J. Lloret, J.P.C. Rodrigues, Smart collaborative mobile system for taking care of disabled and elderly people, Mob. Netw. Appl. 19 (3) (2014) 287–302.
[30] S. Gonzalez-Valenzuela, M. Chen, V.C.M. Leung, Mobility support for health monitoring at home using wearable sensors, IEEE Trans. Inf. Technol. Biomed. 15 (4) (2011) 539–549.
[31] A. Crooks, A. Croitoru, A. Stefanidis, J. Radzikowski, #Earthquake: twitter as a distributed sensor system, Trans. GIS 17 (2013) 124–147.
[32] X. Song, Q. Zhang, Y. Sekimoto, T. Horanont, S. Ueyama, R. Shibasaki, Intelligent system for human behavior analysis and reasoning following large-scale disasters, IEEE Intell. Syst. 28 (4) (2013) 35–42.
[33] T. Horanont, A. Witayangkurn, Y. Sekimoto, R. Shibasaki, Large-scale Auto-GPS analysis for discerning behavior change during crisis, IEEE Intell. Syst. 28 (4) (2013) 26–34.
[34] X. Ma, H. Yu, Y. Wang, Y. Wang, Large-scale transportation network congestion evolution prediction using deep learning theory, PLoS ONE 10 (3) (2015) 1–17.
[35] Y. Lv, Y. Duan, W. Kang, Z. Li, F.-Y. Wang, Traffic flow prediction with big data: a deep learning approach, IEEE Trans. Intell. Transp. Syst. 16 (2) (2015) 865–873.
[36] D. Zhang, T. Hei, Y. Liu, S. Lin, J. Stankovic, A carpooling recommendation system for taxicab services, IEEE Trans. Emerg. Top. Comput. 2 (3) (2014) 254–266.
[37] M. Newman, Networks: An Introduction, Oxford Univ. Press, 2010.
[38] http://www.intel.com.au/content/dam/www/public/us/en/documents/white-papers/big-data-visualization-turning-big-data-into-big-insights.pdf.
[39] C. Chen, J. Hu, Q. Meng, Y. Zhang, Short-time traffic flow prediction with ARIMA–GARCH model, in: Proc. IEEE Intelligent Vehicles Symposium, 2011, pp. 607–612.
[40] M.V.D. Voort, M. Dougherty, S. Watson, Combining Kohonen maps with ARIMA time series models to forecast traffic flow, Transp. Res., Part C, Emerg. Technol. 4 (5) (1996) 307–318.
[41] L. Zhang, Q. Liu, W. Yang, N. Wei, D. Dong, An improved k-nearest neighbor model for short-term traffic flow prediction, Proc., Soc. Behav. Sci. 96 (2013) 653–662.
[42] A. Pascale, M. Nicoli, Adaptive Bayesian network for traffic flow prediction, in: Proc. IEEE Statistical Signal Processing Workshop, 2011, pp. 177–180.
[43] E.I. Vlahogianni, M.G. Karlaftis, J.C. Golias, Optimized and meta-optimized neural networks for short-term traffic flow prediction: a genetic approach, Transp. Res., Part C, Emerg. Technol. 13 (3) (2005) 211–234.
[44] B.L. Smith, M.J. Demetsky, Short-term traffic flow prediction models – a comparison of neural network and nonparametric regression approaches, in: Proc. IEEE Int. Conf. Humans, Information and Technology, 1994, pp. 1706–1709.
[45] C. Chen, K. Petty, A. Skabardonis, P. Varaiya, Z. Jia, Freeway performance measurement system: mining loop detector data, in: Proc. 80th TRB Annual Meeting, 2001, pp. 1–20.
[46] S. Li, L.D. Xu, X. Wang, Compressed sensing signal and data acquisition in wireless sensor networks and Internet of Things, IEEE Trans. Ind. Inform. 9 (4) (2013) 2177–2186.
[47] W. Chen, I.J. Wassell, Energy efficient signal acquisition in wireless sensor networks: a compressive sensing framework, in: Proc. 6th International Symposium on Wireless and Pervasive Computing, 2011, pp. 1–6.
[48] H. Wang, H.E. Roman, L. Yuan, Y. Huang, R. Wang, Connectivity, coverage and power consumption in large-scale wireless sensor networks, Comput. Netw. 75 (2014) 212–225.
[49] Y.L. Tan, V. Sehgal, H.H. Shahri, SensoClean: handling noisy and incomplete data in sensor networks using modeling, Technical Report, University of Maryland, 2005.
[50] R. Zhang, P. Ji, D. Mylaraswamy, M. Srivastava, S. Zahedi, Cooperative sensor anomaly detection using global information, Tsinghua Sci. Technol. 18 (3) (2013) 209–219.
[51] L. Xiang, J. Luo, C. Rosenberg, Compressed data aggregation: energy-efficient and high-fidelity data collection, IEEE/ACM Trans. Netw. 21 (6) (2013) 1722–1735.
[52] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. Le, M. Mao, A. Senior, P. Tucker, K. Yang, A. Ng, Large scale distributed deep networks, in: Proc. Advances in Neural Information Processing Systems, 2012.
[53] R. Polikar, Ensemble based systems in decision making, IEEE Circuits Syst. Mag. 6 (3) (2006) 21–45.
[54] http://www.tableau.com/.
[55] http://www.cytoscape.org/.
[56] H.V. Jagadish, Big data and science: myths and reality, Big Data Res. 2 (2) (2015) 49–52.
[57] K. Kannan, B. Srivastava, R.U. Sosa, R.J. Schloss, X. Liu, SemEnAl: using semantics for accelerating environmental analytical model discovery, in: Proc. Third Int. Conf. Big Data Analytics, 2014, pp. 95–113.
[58] H. Ma, X. Zhang, L. Fan, X. Yan, H. Chen, Y. Gao, The role of big data in regional low-carbon management: a case in China, in: Proc. International Conference on Information Systems, 2014.
[59] V. Vimarlund, S. Wass, Big data, smart homes and ambient assisted living, Yearbook Medical Informatics, 2014, pp. 143–149.
[60] S.J. Redmond, N.H. Lovell, G.Z. Yang, A. Horsch, P. Lukowicz, L. Murrugarra, M. Marschollek, What does big data mean for wearable sensor systems?, Yearbook Medical Informatics, 2014, pp. 135–142.
[61] N. Asadi, J. Lin, Fast candidate generation for real-time tweet search with bloom filter chains, ACM Trans. Inf. Syst. 31 (3) (2013).
[62] B. Abedin, A. Babar, A. Abbasi, Characterization of the use of social media in natural disasters: a systematic review, in: Proc. IEEE Fourth Int. Conf. Big Data and Cloud Computing, 2014, pp. 449–454.
[63] N. Kumar, S. Misra, J.P.C. Rodrigues, M.S. Obaidat, Coalition games for spatio-temporal big data analytics in Internet-of-Vehicles environment: a comparative analysis, IEEE Int. Things J. (2015).
[64] S. Djahel, R. Doolan, G.-M. Muntean, J. Murphy, A communications-oriented perspective on traffic managements systems for smart cities: challenges and innovative approaches, IEEE Commun. Surv. Tutor. 17 (1) (2015) 125–151.
[65] C.-Y. Chong, S.P. Kumar, Sensor networks: evolution, opportunities, and challenges, Proc. IEEE 91 (8) (2003) 1247–1256.
[66] W. Ye, F. Silva, A. DeSchon, S. Bhatt, Architecture of a satellite-based sensor network for environmental observation, in: Proc. Earth Science Tech. Conf., 2008.
[67] M.I. Poulakis, S. Vassaki, A.D. Panagopoulos, Satellite-based wireless sensor networks: radio communication link design, in: Proc. 7th European Conf. Antennas and Propagation, 2013, pp. 2620–2624.
[68] C. Albaladejo, P. Sanchez, A. Iborra, F. Soto, J.A. Lopez, R. Torres, Wireless sensor networks for oceanographic monitoring: a systematic review, Sensors 10 (7) (2010) 6948–6968.
[69] H. Yang, H. Wu, Y. He, Architecture of wireless sensor network for monitoring aquatic environment of marine shellfish, in: Proc. 7th IEEE Asian Control Conf., 2009, pp. 1147–1151.
[70] S.J. Pan, D. Shen, Q. Yang, J.T. Kwok, Transferring localization models across space, in: Proc. Twenty-Third AAAI Conf. Artificial Intelligence, 2008, pp. 1383–1388.
[71] V.W. Zheng, D.H. Hu, Q. Yang, Cross-domain activity recognition, in: Proc. 11th Int. Conf. Ubiquitous Computing, 2009, pp. 61–70.
[72] P. Rashidi, D.J. Cook, Activity recognition based on home to home transfer learning, in: Proc. 24th AAAI Conf. Artificial Intelligence, 2010, pp. 45–52.
[73] P.K. Atrey, M.A. Hossain, A.E. Saddik, M.S. Kankanhalli, Multimodal fusion for multimedia analysis: a survey, Multimed. Syst. 16 (2010) 345–379.
[74] G. Jaffre, J. Pinquier, Audio/video fusion: a preprocessing step for multimodal person identification, in: Proc. Int. Workshop Multimodal User Authentication, 2006.