IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 16, 2018, accepted October 3, 2018, date of publication October 8, 2018, date of current version October 31, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2874500

Spatial Big Data and Moving Objects: A Comprehensive Survey

USAMA MIR[1], UBAID ABBASI[2], YANG YANG[3], (Member, IEEE), ZEESHAN AHMED BHATTI[4], AND TALHA MIR[5]

[1]Saudi Electronic University, Saudi Arabia
[2]GPRC, Grande Prairie, AB T8V 4C4, Canada
[3]Center for Data Science, Beijing University of Posts and Telecommunications, Beijing 100876, China
[4]King Abdulaziz University, Jeddah 21589, Saudi Arabia
[5]Tsinghua University, Beijing 100084, China

Corresponding author: Usama Mir (u.mir@seu.edu.sa)

This work was supported in part by the National Science Foundation of China under Grant 61801035 and in part by the 2018 Education Basic Project under Grant 500418763.

ABSTRACT Recently, big data (BD) has seen a tremendous growth in its volume, magnitude, and complexity. Examples of such data include navigation maps, mobile phone trajectories, and social media posts/tweets. Normally, this humongous data is known as spatial BD (SBD). The handling and routing of SBD datasets become more challenging when the movement of objects (such as humans and vehicles) is highly dynamic and random. In this paper, we focus on different aspects related to generation, routing, and handling of BD and SBD. The purpose of this paper is multifold; first, we present the viewpoint of various researchers about BD. This part also includes the differentiation between BD and SBD based on several examples. Second, we focus on social media and e-applications which are considered to be the biggest contributors in generating large volumes of spatial data. Third, this paper highlights the routing perspective for BD and SBD including various interesting strategies to route large traffic volumes generated by moving objects. Fourth, we discuss different techniques for big data analysis within the context of moving objects. Finally, we highlight important issues and challenges within the domain of BD and SBD.

INDEX TERMS Big data, moving objects.

I. INTRODUCTION

The enormous spread of low cost sensing devices and dramatic drop in storage cost of the data has resulted in extraction of valuable information that is used for the efficient design of any system. The volume, velocity, and variety of such big data (BD) generation and processing is unprecedented. The benefits of BD processing can be seen in a variety of applications such as device to device (D2D) networks [42], urban planning [161], traffic safety [162] and Intelligent transport system (ITS) [90].

There are various sources that contribute to the generation of BD ranging from social media, global positioning maps, electronic and marketing apps (including e-marketing and e-health), global climate models, and vehicle-related information. The importance and handling of such BD become more significant and complex if the objects (generating data) are highly mobile and spatially located at a distance. This type of data is mostly referred to as spatial big data (SBD) [13], [90].

One of the biggest challenges for SBD is the dynamic nature of spatially distributed devices and objects especially with the availability of 4G-LTE (long term evolution) services around the globe [113]. The limitation of spectral efficiency for the 4G physical layer results in the utilization of highly dynamic and resource efficient networks such as D2D [42]. In addition, the presence of navigation services such as Google Maps has started a new era of GPS (global positioning system)-based communications. Therefore, this highly dynamic nature of data requires us to develop real-time solutions in order to track the moving objects [90].

Another vital issue is to efficiently 'route' the SBD [96]. This means routing the traffic patterns generated by the movements of objects (such as vehicles and humans) from one location to another. Such kind of information is highly random and quite difficult to manage. Another perspective on routing is to handle the amount of big data generated by modern day networks such as clouds and software defined networks (SDNs) [64].

In addition, an important phenomenon for BD and SBD is extracting the appropriate conclusion by properly analyzing the collected data. The exploitation of collected data has always received considerable attention by academia and industry, however, huge volume, high velocity, and large variety of BD push the limits of the existing storage and processing systems. Moreover, it is difficult to apply the existing statistical techniques on this huge data due to several reasons such as the biasness. The collected data consists of both structured and unstructured information that might be useful for a specific scenario only [25].

Therefore, it is highly important to develop modern techniques and methods that assist in processing these huge data streams. There are different dimensions of research in the area of BD, but in case of BD analysis for moving objects, it mainly relies on ad-hoc solutions [23], [51]. The emphasis stays on achieving certain predefined objectives (such as traffic flow predictions, mining data from sensors, etc.) which may result in very limited applications and range of scenarios. Currently, there are variety of sources for data generation in moving objects including social media, video devices, sensors, and GPS, however, it is not enough to rely only on these sources for extremely sensitive decisions such as real-time traffic flow, routing, and ITS. In addition, the new trend in moving objects is more towards proactive control such as accurate prediction of congestion or traffic accidents. Therefore, it is crucial to understand the inherent mechanism of moving objects using both historical as well as real data.

A. RELATED SURVEYS AND CONTRIBUTION OF THIS ARTICLE

In relation to above, a few survey related efforts have been done in the recent past. For example, the work proposed in [90] briefly differentiates between BD and SBD. This paper explains in detail the basic concept of big data and differentiates it with SBD using interesting examples. Several types of big data are also defined such as raster, vector, and graph based datasets. Later, Michael *et al.* [90] shift their entire focus on cyber-based information which includes fault-tolerant file systems and BD clouds. Thus, the main idea of interest roams around global information systems (GIS). Our work, on the other hand, not only covers the spatial aspect of big data but also highlights the social media, routing, and learning perspectives. We list down numerous definitions related to SBD, and explain in detail various approaches and algorithms for SBD generation, handling, routing, and analysis. Another short survey is presented in [177] highlighting important challenges and opportunities in trajectory data analysis. The paper is written on the argument that current research on SBD trajectory analysis focuses on a single factor (such as the movement of people from one place to another) which is quite simple and thus, not enough to make effective and real-time routing decisions. Therefore, the authors emphasize on including multiple factors in trajectory analysis such the types of vehicles used for movement, temperature, and road conditions and they list challenges and issues related to the aforementioned. However, we think that the claim of authors is not totally justified since there are several existing approaches (such as [30] and [90]) which consider multiple factors for routing-based decisions.

The work proposed in [176] highlights the challenges for SBD with uncertainties. The focus of the paper is to identify the unreliable knowledge resulting from different sources and ensure stage-by-stage uncertainty handling for enhancing the reliability of the knowledge. The authors explain the reasons for the uncertain data and provide a framework for uncertainty based SBD analytics. They further discuss the place-based heuristic, analytics, and assessment of user generated images and texts in the context of any existing uncertainty. Quite differently, our work (especially section V) highlights different aspects of big data analysis in moving objects. Although uncertainty can be classified as one of the issues in SBD analysis for moving objects however, we kept our focus on discussing existing techniques used for effective analysis of the available data. Another interesting survey presented in [79] lists the important platforms for BD analysis but the importance of spatiality and moving objects is mostly ignored. The focus of [50] remains on disaster management aspects related to BD networks. The networking of BD is very comprehensively covered in [107]. Xia *et al.* [28] survey all the issues related to the processing and handling of BD which is generated via scholarly information databases. In this paper, we focus on various aspects related to SBD and moving objects. More specifically, the main contributions of this paper are as follows:

- We comprehensively define BD. This part includes the viewpoint of several researchers about BD. BD is then differentiated with SBD with the help of promising examples. This section also clarifies the relationship between SBD and moving objects.
- We focus on social media and e-applications which are considered to be the highest contributors in generating large volumes of spatial data. This section also includes examples of current market trends related to BD.
- Our study highlights the routing perspective for BD and SBD including various interesting strategies to route large traffic volumes generated by the moving objects. This part surveys different strategies designed for routing of vehicles as well as network-generated traffic.
- We discuss different techniques for BD analysis within the context of moving objects.
- Current and future challenges and issues are also discussed for all of the above mentioned parts.

The acronyms used throughout this article are presented in Table 1.

B. SURVEY STRUCTURE

The paper is organized as follows. In the next section, we comprehensively define BD, differentiate it with SBD using examples, and explain its relationship with moving objects. Important contributing factors involved in SBD generation

TABLE 1. List of acronyms/abbreviations.

Acronym/Abbreviation	Full form
ADP	Adaptive Dynamic Programming
ANBR	Aggregatable Named Based Routing
AODV	Ad hoc On demand Distance Vector
API	Application Programming Interface
ARM	Association Rule Mining
AV	Autonomous Vehicle
BD	Big Data
BS	Base Station
CCN	Content Centric Network
CR	Cognitive Radio
CRD2D	Cognitive Radio Device to Device
D2D	Device to Device
DACAR	Data Capture and Auto identification Reference
DCN	Data Center Networks
DSR	Dynamic Source Routing
EAR	Explicit Adaptation Requests
GIS	Global Information System
GPS	Global Positioning System
IaaS	Infrastructure as a Service
ITS	Intelligent Transport System
LTE	Long Term Evolution
MAE	Mean Absolute Error
MANET	Mobile Ad hoc Network
MAP	Mean Average Precision
NS2	Network Simulator 2
PDR	Pick-up and Drop-off Rate
PPCA	Probabilistic Principal Component Analysis
PU	Primary User
RFID	Radio Frequency Identification
RL	Reinforcement Learning
ROI	Return On Investment
RST	Rough Set Theory
SA	Sentiment Analysis
SaaS	Software as a Service
SBD	Spatial Big Data
SDN	Software Defined Network
SU	Secondary User
TOU	Time of Use
V2V	Vehicle to Vehicle
VGI	Volunteered Geographical Information
WWW	World Wide Web

(such as social media) w.r.t moving objects are highlighted in Section III. Routing is directly related to movement of spatially located objects, therefore Section IV entirely focuses on various routing strategies for BD and SBD. BD analysis techniques are discussed in Section V. We highlight current and future research issues in section VI. Finally, Section VII concludes the paper.

II. SPATIAL BIG DATA AND MOVING OBJECTS

The term big data signifies to huge amount of data that is complex to analyze. There are several examples of BD such as location maps, investigative reports, data gathered from thousands of sensors placed at various locations, climatic changes and detailed weather reports, political surveys about upcoming elections, and so on. According to [99], big data is large and heterogeneous making it difficult to store in regular relational databases. This data can either be structured (stored in relational databases) or unstructured to be used for analysis.

Michael *et al.* [90] argue that the size of data varies with the context which means there is a solid correlation between the dataset and users' personal experiences. Snijders *et al.* [15] define BD as a combination of datasets which are humongous in size and cannot be solved using simple statistical techniques within the allocated time frame. These datasets are mostly loosely coupled showing the complexity in handling and processing. Similarly, Laney [21] defines BD as any information with large size, variety, and high volume that can only be managed through the development of advanced optimization and discovery techniques. Moreover, well-known organizations like IBM, Intel, and Oracle also distinguish BD from conventional data in its size, volume, and unstructured nature [47]. Table 2 summarizes various existing viewpoints about BD.

BD is in limelight for years now and a lot of research has been done on defining, processing, and handling these large sets of random information. Due to the current advances in the field of information technology especially with the provision of internet connectivity whenever and wherever needed, SBD has emerged as one of the most important and interesting research areas. The terms BD and SBD are easy to differentiate. For example, normal posts/tweets on Facebook and Twitter can be categorized as BD while geo-located posts/tweets from different parts of the world fall under the window of SBD [90]. Another example of SBD is the data gathered from various 'sensor nodes' spatially located in a large geographic region. This data has to be stored, monitored, and processed in real-time to deal with situations like disaster recovery, possibility of fire spreading in a jungle, and temperature control.

SBD is largely influenced by the movement of geographically located objects. These 'moving objects' are mostly in the form of humans and vehicles changing their locations on frequent basis. Sometimes, these location changes are so random and fast that it gets difficult to process and control the generated data. Thus, special algorithms and solutions are required to handle the data generated by the moving objects.

Recent trends shown by the mobile industry in terms of volume (up to Terabytes) and variety (such as gaming, multimedia, cloud services, navigation, and social posts) are also part of SBD. A huge increase in the usage of location-based services (like Google Earth, Google Maps, Swarm, Waze, etc.) through mobile phones and other handheld devices highlights

TABLE 2. Viewpoints of researchers from academia and industry related to big data. These viewpoints are chronologically arranged according to the year (and date in case of same year) of publication.

Type of published document	Concept about Big Data	Reference/ Source	Year of Publication	Context
Research article	BD is defined based on the volume and size of the data.	[66]	1997	Computer graphics
Technical report	Amount of data which is difficult to manage by the modern day database software.	[46]	2011	Enterprises storage and consumers usage
Research article	BD is defined in terms of its size and rapid generating nature.	[101]	2012	Computing
Study report	A large volume of data which is generated by new applications and technologies and thus, requires advanced processing techniques.	[40]	2012	Digital communication
Book	BD is defined according to culture and society values and people's ability to use information.	[116]	2013	Things that can be done with big data (Best seller in USA)
Research article	BD is humungous and requires advance level processing techniques.	[46]	2013	Computing
Research article	Similar idea as [40] and [46].	[99]	2013	General (not specific to a particular topic)
Research article	Kind of data which requires timely processing however modern day tools are finding it difficult to process BD in time.	[112]	2013	Computing
Book	Large datasets seem to emerge from last decade or so.	[110]	2014	Management sciences
Online page (blog)	Large variety and volume of datasets requiring innovative techniques for processing.	[9]	N/A	General
Research article	BD is defined based on three 'C' i.e., *Cardinality, Continuity,* and *Complexity*.	[105]	2014	Machine learning
Technical report	Extensive datasets which require scalable processing techniques.	[24]	2015	General

the importance of SBD in our daily life. Table 3 summarizes the difference between BD and SBD w.r.t several interesting contexts and examples.

In continuation to above, figure 1 depicts a case study based on people's daily patterns in life. These statistics were presented in [103] where the 'activity space' is chosen as a life pattern. Based on a thorough GPS tracking of 30 people's weekday routine, we see that a high percentage spends their time at homes or in offices. However, their online data access could totally be different. For example, from noon to 5pm, the pattern of data would be video conferencing, stock exchange monitoring, fund transferring, tweeting about current achievements in office, and posting pictures of lunch menus and restaurants. On the other hand, the hours from midnight to 7am would result in rather different traffic patterns like normal web-browsing and social media activities. Therefore, a person having a seamless internet connectivity will be experiencing a wide variety of activity during a single day and by multiplying it with billions, we can think of how huge, dynamic, and geographically distant this data can become. Furthermore, in this context, in Table 4, we name certain examples of data, which are specific to SBD and moving objects, and should be processed in real-time. The recent trends show that the total amount of data (generated from the types specified in Table 4 and similar other cases) is increasing at a fast pace and expected to reach 7.9 Zettabytes this year compared to 1.8 Zettabytes in 2011, respectively [84]. Therefore, efficient techniques are desirable to compensate this immense growth in SBD considering the distributed and dynamic nature of objects.

III. MAIN SOURCES FOR BIG DATA GENERATION: SOCIAL MEDIA AND E-AAPLICATIONS

One of the main sources of BD is the extensive use of social media applications as shown in Table 4. The substantial growth of social media over the last decade has been

TABLE 3. A brief comparison of big data and spatial big data considering different contexts. Some well-known example applications for SBD are also provided.

Context	Big Data	Spatial Big Data	SBD Examples
Social Media	Simple posts or tweets	Geo-located posts or tweets	Facebook, Twitter, Whatsapp
Searching	Menus of all Chinese restaurants in the middle east	Menus of all restaurants in the middle east with more than 100 hundred guests check-ins during last couple of hours	Foursquare
Surveillance and Monitoring	Detailed study manuals on deployment of sensor nodes and security cameras	Change of sensory data in last hour or so in all the places where temperature sensors are deployed within a country	Loop detectors, cameras, hotspots
Political Analysis	Results of previous two elections	Voting trends in past 48 hours	Buzzfeed
Television/Movies	Checking www.imdb.com	Detailed channel ratings based on viewers presence during particular hours	USA NOW
Navigation	Detailed past history of all reported accidents on the highway connecting eastern and western regions	Real-time maps with shortest paths, nearby gas stations, restaurants, and hospitals	Google Maps
Daily Life	Detailed biography	Patterns of life for last three months including work, sports, club, study, sleeping hours, and surfing on mobile/laptop	UBhind

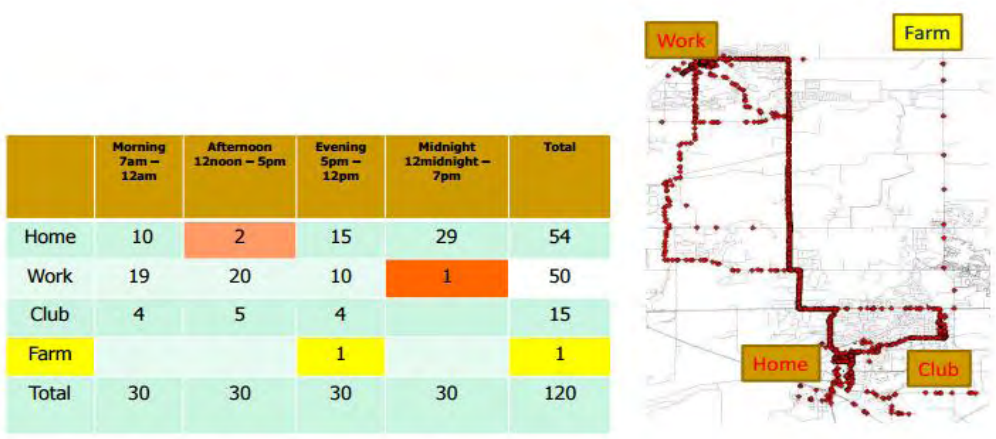

FIGURE 1. A use case study of people's weekday activity patterns. Study is conducted for a period of three months by tracking (through GPS) the continuous spatial movements of people from one activity space to another [103].

supplemented by the upsurge of spatial technologies. This has provided new mapping mechanisms where users are involved in online activities in real time. The coupling of users social media and location-based information generates tons of crowdsourced data that can be used in many contexts. However, this unstructured data requires advanced data mining techniques to draw meaningful insights. Social media data together with geotagging, known as geosocial data enables content creation by users and is publicly accessible. Thus, social media is one of the main sources of SBD generation. This is mainly because of the exponential growth of mobile usage thus making it possible to measure

TABLE 4. Types of modern day real-time services generating huge amount of data. This data is strongly related to the context of moving objects.

Type of Real-time Data	Famous Example(s)
Social Network Data (also mentioned in Table 3)	Facebook, Twitter, Instagram
Click Streams Data	YouTube, Dailymotion, Broadcast through privately own smart-phones (such as Samsung Note)
Automobiles Data	Google Maps, TomTom, Samsung Mobile Live Broadcast
RFID (radio frequency identifier) Data	Smart cards, goods tracking
Location-Based Data	GPS, Whatsapp location feature. Also includes automobile data
Point of Sale Data	Barcodes
Customer Data	Billing, Inquiries, Payments
Smart Meters Data	Digital electricity meters for TOU (time of use) tariffs and fluctuation management
Sensory Data	Security cameras, temperature and climate sensors
Mobile Data	4G-5G and LTE services

and analyze different types of social, economic, and political activities [131]. BD generated via social media is also known as 'social BD'. This social BD is defined in [132] as "*those processes and methods that are designed to provide sensitive and relevant knowledge to any user or company from social media data sources when data sources can be characterized by their different formats and contents, their very large size, and the online or streamed generation of information*". One of the features of data generated through social media applications is that it is unstructured in nature. Due to the various types of social media generated data, the analytics performed on social BD vary as well. Text analytics are used for social media data collected from organizational proprietary web pages as well as pages hosted on various third party social media websites. Social media content helps firms in understanding the opinions of consumers and general public related to social events, political movements, company strategies, marketing campaigns, and product preferences [133]. In order to better understand these opinions, sentiment analysis (SA) techniques are used in literature. SA is a computation study of opinions, sentiments, emotions, and attitudes expressed in text towards an entity [134]. Due to the ever increasing trend in e-commerce, online reviews and product evaluations are becoming common, and therefore, producers and service providers take into consideration the analysis of these customer opinions [135]. However, the increasing use of mobile devices coupled with social media data has provided organizations an access to much more in-depth data which is even more useful. This is usually referred as 'mobile analytics' or 'mobile business intelligence'. According to Chen *et al.* [136], mobile data generation and analytics have a wide range of applications such as location-based mobile sensing apps that are activity-sensitive, disaster management, m-health, m-learning, mobile social networking, crowd-sourcing, gamification, mobile advertising, and social marketing. Some of these applications are discussed in detail in the following subsections.

A. MOBILE/SOCIAL MARKETING

Marketing activities with social BD have been enhanced to a great extent. Organizations are investing in BD infrastructure by logging and storing data from customer engagements because it helps to achieve a positive return on investment (ROI) [137]. One way to get high ROI is to perform predictive analytics for product demands which offers accurate demand forecast. According to [137], traditional time-series forecasting with historical sales data works well with popular products but is not very predictive for other products because of the random noise, hence more and more firms are relying on predictive analytics from socially generated consumers data.

Online social networking websites provide marketers with deep insights and user intelligence to launch targeted marketing campaigns. Similarly, online recommender systems used by YouTube and Netflix as well as e-commerce websites such as Amazon and eBay can significantly transform marketing [132]. A study of Twitter as an advertising medium analyzing range frequency, timing, and tweet contents revealed that 19% of tweets from the corporate accounts do mention at least one 'brand' [163].

He *et al.* [145] argue that in today's competitive environment, BD is one of the vital resources for a firm that is able to generate something of value to its customers and cannot be imitated. According to the resource-based view of a firm, there are three types of resources: physical capital, human capital, and organizational capital [139]. In the context of BD, the physical capital is referred as the applications used to gather and analyze BD [164]. The human capital can be equated to the data analysts who can have insights into the data to know about consumer activities. The organizational capital is basically the structure that enables the strategies to be translated into actions [138]. Building on this insight, firms have recently started utilizing consumer BD generated through mobile phones and applications to enhance their dynamic capabilities. One such example is the use of recorded conversations between service staff and customers by

various organizations with the integration of speech-analytics to determine key performance indicators [140]. Similarly, a much more proactive approach is used by Target stores in competition with Walmart reported by Li *et al.* [142].

With the built-in GPS in mobile devices to enable location information for users, social media applications allow their users to post content with geographic information, hence producing vast amount of data with location information [168]. Gonzalez *et al.* [169] coined the term individual mobility pattern (IMP) to understand human behaviors in performing daily activities based on their locations. The main source of data for IMP which is mobile phone trajectory combined with social media data generated by users becomes a competitive data source for moving objects. Naturally, with respect to IMP, spatial aspect remains the most important one e.g., frequent visits to point-of-interest as identified by Yuan *et al.* [170] and the role of GIS as social media by online mapping websites. Therefore, a plenty of research is done on social media and spatial data. However, researchers have recently realized regularities not only in spatial but also in temporal aspects of IMP and hence stressing spatial-temporal data for individuals to be more rational [168]. Therefore, in addition to the data related to individuals points-of-interest, temporal data such as the sequence of visits to these points-of-interest provides us with more comprehensive spatio-temporal data.

He *et al.* [145] state that with the growing richness of data, marketers now have more opportunities to identify new gaps which was not possible before. With the availability of geospatial and temporal data, marketers can now predict where a customer/product would possibly be at a given time. Google Now application has the capacity to suggest consumers on the availability of products/services based on their browsing and searching patterns e.g. availability of tickets at a local cinema and movie hours (based on temporal data) or a restaurant offering food according to customer's likings.

Text Analytics and Network Analytics in Social Media: One source of SBD (particularly for organizations) is the unstructured textual information residing in email communications, corporate documents, and social media content on websites such as Facebook, Twitter and other social media. Government and research institutions also monitor such textual information for security purposes and/or to analyze the general opinion of public for popular events such as elections and referendums.

According to Erevelles *et al.* [138], in the World Wide Web (WWW), opinion identification and classification is normally based on social media textual contexts, therefore the authors suggest two main approaches for such classification; lexicon and learning based methods. Lexicon-based approaches have pre-established rules to interpret textual sentiments [143] such as measuring the tone of news. Learning methods (on the other hand) use a hybrid approach combining lexicons and machine learning to classify opinions (the learning methods are discussed in much detail in section V). Similarly, online reviews of products by people/customers are also a source of data for marketers as well as producers [144]. Qi *et al.* [144] further argue that online reviews allow a customer to be a part of product design and development as well as customer requirements.

Erevelles *et al.* [138] believe that opinion mining is an extensively used approach in social media environments. Network analytics are often coupled with text analytics in social media to derive more sense out of data. Users' online posts or tweets trigger responses such as likes, comments, and re-shares, which may lead to the population of more textual content. According to Chen *et al.* [136], network analytics focus on link mining and community detection by exploring and predicting the connectivity among nodes of a network. Nodes can be in the form of customers and users of a service or products, while the links represent the medium such as email messages, and social media comments, likes and shares. These connections between nodes create a social network of opinionated users.

Online recommendation system such as YouTube video recommendation also utilizes SBD consolidated by network connections by grouping similar users and extending it to other users by suggesting content they might be interested in. For instance, He *et al.* [145] show that an online service for travel recommendation must capture or infer users' travel interests from their consumption behavior on travel packages which is termed as social recommendation.

Network SBD analysis is also being utilized in road network modeling for instance in an online taxi service with the use of GPS sensors. Zhou *et al.* [146] highlight a problem in traditional taxi service focusing on misbehavior of taxi drivers. These drivers were not taking the shortest route to travel from point A to B, and the passengers were not aware of this deception especially when they were not acquainted with the route. This detouring allows taxi drivers to overcharge passengers in unfamiliar settings. By incorporating GPS sensors in taxis, data such as position, speed, timestamps, and flag markings can be generated. Therefore, Zhou *et al.* [146] suggest to prevent this fraud by considering the data of customers who have already traveled from point A to B and hence, consolidating this data to calculate the optimal route. Zhou *et al.* [146] further recommend a real-time trajectory detection which means that when a driver tries to deviate the recommended path, a notification may be pushed to the passenger notifying him/her of a potential fraud. We further discuss route recommendation and similar topics in Section IV.

B. M-HEALTH

The extensive use of e-health records by hospitals across the globe has generated vast datasets [147]. Apart from hospitals, there are other sources of BD in healthcare such as clinical decision support systems, laboratories, pharmacies, and insurance companies. The gathering and analysis of healthcare related BD has been given various terminologies such as e-health, m-health (medical-health), digital health, health 2.0, and e-medicine [148]. One of the

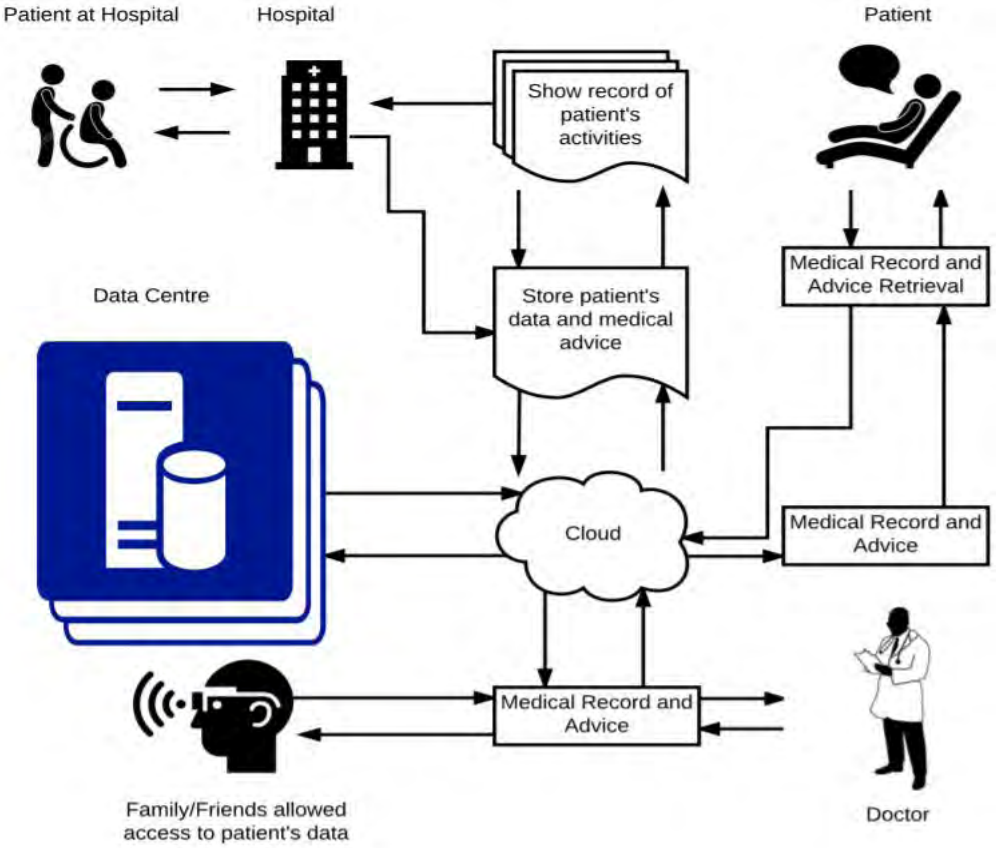

FIGURE 2. M-Health System at Chelsea and Westminster Hospital (London) (Adapted from [153]).

features of such data is crowdsourcing which is the basis of social media technologies. A number of websites has emerged recently trying to consolidate crowdsourced data such as PatientsLikeMe, Health Tracking Network [149], and Global Public Health Intelligence Network [132]. Barrett *et al.* [149] state that crowdsourced data is especially beneficial for observing the spread of infectious diseases. Similarly, HealthMap – an online aggregator uses informal online data sources such as news, reports, and online discussion forums to observe and map diseases-related information. In the same fashion, individual users' (Facebook and Twitter) profiles generate diseases-related data using the keywords in their status updates/tweets which helps in mapping the location and intensity of a disease.

In addition to above, a more accurate source of such data is the use of mhealth applications installed on smartphones and sensory wearable devices [150]. The growing trend of health related applications on smartphone has triggered the market for personal sensors with pre-loaded applications – which serve as passive and manual data generation source [149]. In addition, devices such as Google Glass, Apple iWatch, and Samsung Gear among others have built in hardware and software functionalities for monitoring users' heart, pulse, and respiratory rates [151]. Sultan [152] reported the life-saving incident of a patient using Google Glass at Boston's Beth Israel Deaconess Medical center where physicians learned a patient's complete medical history from his wearable device. This patient had earlier provided medical staff with an incomplete medical history. Google Glass helped the physicians in an emergency situation and saved time accessing patient's medical records with more accuracy than searching through physical records. Following this incident, the hospital extended the use of Google to the entire emergency department.

Another potential use of wearable devices in healthcare is for data access. Sultan [153] mentioned the use of DACAR (data capture and auto identification reference) to gather remote data from patients that allows close family and friends to access data buckets of patients for surveillance and care purposes. Family members or friends can set threshold values (e.g., heart rate and blood pressure) in the system to trigger alarms related to a patient's health condition. Figure 2 shows the functioning of the m-health system 'DACAR' adapted by the Chelsea and Westminster hospital in London [147]. The system is basically a platform that enables remote delivery, storage and access to patient's electronic records using platform as a service (PaaS). DACAR is hosted on an IaaS (infrastructure as a service) cloud platform. The system enables close friends and family of a patient as well as the hospital/doctors to access his/her data buckets at

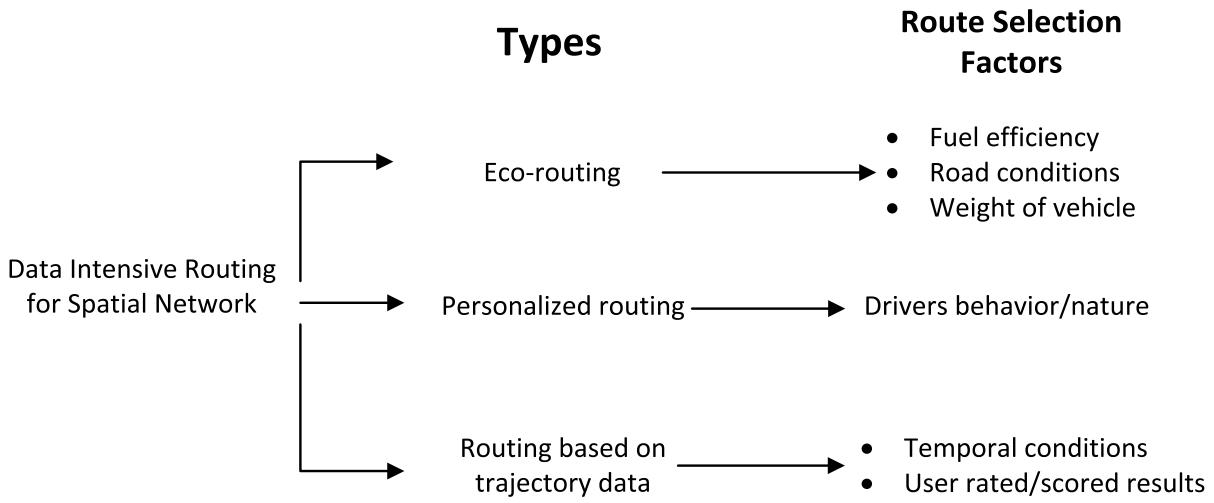

FIGURE 3. Types of data intensive routing with route selection factors for each type [14].

any given time. The system is essentially connected to wearable technology on a patient and a data bucket allows creating, holding, updating, and deletion of patient's data. A cloud server stores all the updated data which is accessible by multiple stakeholders such as friends/family for monitoring the patient, the patient itself, and the doctor for accessing medical records. A consulting doctor or a staff can always access the real-time updates of any patient, as the data is continuously saved on the cloud. Data from the cloud can also be sent to the data center for further analysis. In addition, the DACAR system allows a patient to easily share information with the consultant if he/she is to seek any medical advice. Finally, for critical and elderly patients, their close family members are notified via messages if any health monitoring values breach the threshold levels (such as high blood pressure, pulse, and heartbeat).

C. SBD GENERATED THROUGH DISASTER SITUATIONS

Recent natural environmental disasters have drawn focus to the vulnerability of our society and infrastructure [154]. Novel information streams, such as social media-contributed videos, photographs, and text as well as other open sources are redefining situation awareness during emergencies [155]. Social media messages combined with geographic reference containing spatial and temporal information are termed as volunteered geographical information (VGI) [156] that unfolds the power of 'citizens as sensors' to provide real-time updated data. According to Dashti *et al.* [157], VGI can be used in disaster response, timely damage assessment, and to efficiently carry out rescue and relief operations. Howe [158] argues that VGI is similar to crowdsourcing primarily because of two reasons. First, it makes use of the wisdom of crowd and can solve a problem more effectively than an expert, irrespective of the group's knowledge and expertise pertinent to the problem. Second, the information acquired via collaborative VGI is possibly closer to reality than the information obtained from a single source. A number of examples where volunteered data has been used to retrieve critical information have been cited in the recent literature, e.g. during hurricane Katrina [155], fire emergency event in Shanghai [159], River Elbe flood in Germany [160] and so forth.

IV. ROUTING AND BIG DATA

In this section, we summarize several existing contributions related to geographically generated BD. In the first part, we focus on existing solutions which deal with BD traffic generated through geographical/spatial movements of various objects including cars, taxis, and buses. In the second part of this section, we summarize interesting algorithms proposed for the routing of BD traffic in modern day networks such as SDN (software define networking), D2D networks, BD clouds, etc.

A. ROUTING FOR SPATIAL/GEOGRAPHICAL MOVEMENTS OF OBJECTS

BD can play an important role in developing efficient traffic routing schemes. The increase in number of vehicles using location-based services and uncertainty in people's movement from one location to another demand us to develop more accurate, cost-effective, and energy efficient traffic routing schemes [14]. This requires a thorough and deep analysis of the currently available big volumes of data and sharing the available information with users/devices/vehicles in real-time [108].

1) IMPORTANT TYPES OF SPATIAL BIG DATA ROUTING [14]

In relation to routing of SBD, a promising work is presented in [14], where routing is divided in several interesting categories such as eco-routing [57], [72], personalized routing [37], [83], and vehicle routing with user-generated trajectory data [115], [118] as shown in figure 3. In eco-routing, the best route for vehicles is chosen based on fuel efficiency. Other factors such as road situations, traffic and

weather conditions, and amount of luggage in a vehicle can also be considered. Further, two approaches based on histogram and graph theory to represent various case studies related to eco-routing are presented and several interesting results are shown. Different from eco-routing, personalized routing presented in [14] considers a driver's personal preference as a primary factor in selecting a route. For example, someone traveling for vacations during summers will go for fuel economy since he/she has plenty of time to travel. However, any other person driving in a rush hour would be less concerned with fuel saving and would opt for the fastest/shortest route, thus, showing two different preferences of drivers. The authors further suggest to look for a driver's nature in route selection such as aggressive (or a fast-paced driver) and moderate (or a low -paced driver). Despite highlighting this important techniques for route selection, Jensen *et al.* [14] do not provide any further details. Third type of routing presented in [14] considers temporal conditions (such as familiar/unfamiliar surroundings) and popularity index (rated by various local travelers) in selecting the best route. A ''trajectory database'' is created which provides trip suggestions to drivers in real-time. This database has routes which are rated/scored by various drivers. The results of trajectory database are compared with the results provided by the routing service (such as driver's navigation/GPS) and the route with maximum match is selected. Similar kind of routing concepts are highlighted in [41] and [67] as well.

Based on the above concepts, we categorize the existing literature on big data routing in numerous categories discussed in the next subsections. Our work mostly focuses on (but not limited to) the research work done in the past two to three years.

2) GROUPING OF AUTONOMOUS VEHICLES WITH PRIVACY PRESERVATION [1]

Autonomous vehicles (AVs) have emerged as a new phenomenon in modern era with the flexibility of driving without the need of a driver. These vehicles are normally equipped with advanced sensing and navigation facilities allowing them to route with accuracy. The work proposed in [1] tries to utilize these interesting features of AVs. Basically, for each AV, there is a set of primary and secondary users. A primary user (PU) is the owner of an AV and a secondary user (SU) is the one wanting to share a PU's vehicle. A PU decides whether to share its AV with SU(s) on the basis of nearest neighbor selection algorithm proposed in [78]. A PU also has the right to decide the number of SUs who can share its AV. Moreover, three different routing scenarios are presented such as:

- A PU decides the pick and drop areas for SUs according to the start and the end locations agreed by both primary and secondary users. The vehicle is then routed accordingly.
- A PU decides the pickup area but the dropping location is different from the end location, however it should not be far from the surrounding of route.
- A PU decides the pick area but the dropping location is different from the end location and it can be far from the route. The AV can then drop the PU first, take SU to its destination, and finally route back to the PU (which is unlikely to happen with human driven cars).

All the above routing strategies are tested on real maps of Tennessee State (USA) and the results show that the proposed scheme incurs less route searching time which is highly desirable for modern day users dealing with BD applications and AVs.

3) PERSONALIZED ROUTE RECOMMENDATION BASED ON BIG SOCIAL MEDIA

As mentioned before (in Section III), Social media (such as Facebook, Twitter, and Instagram) is one of the biggest examples of BD [31]. We use social media to post photographs and videos, tag friends, make suggestions, and so forth. Thus, social media has become a necessary part of our daily routine. Among other important things shared on social media these days, most people prefer sharing their travel experiences in the form of photographs and posting status updates. This important source of travel information from social media along with various 'Travelogues' available on several websites can be used together to propose more personalized and efficient travel recommendations to users which is done in a very interesting work presented in [96]. The authors propose a learning model named as ''topical package model (TPC)'' which recommends personalized routes to users based on community-contributed photos and travelogues. In addition, the routes are learned and suggestions are made according to the average amount of money a user spends on a trip plus the visiting season and time. Different mathematical models are proposed to show how important information (such as users preferred season for traveling) can be extracted/mined from various social media photos, tags, and travelogues. Mining information is then used to recommend routes to users based on a ranking system. Extensive set of experiments are performed on real-time travelogues to show the impact of user interests, traveling season, time, and cost on mean average precision (MAP-which is calculated using the proposed mathematical model). This MAP reflects the change in choices of users when they are selecting their traveling packages. The results conclude that time and cost have more impact on MAP than traveling season which is kind of an obvious trend. We summarize the main points of [96] and other important works (similar to [96]) which are based on recommending personalized routes to users using different criteria such as geographical locations, distances, time, cost, etc., in Table 5.

For example, in Table 5, the work proposed in [128] is summarized that uses real-time datasets for the experiments collected from Flickr API. Similar to [96], Shi *et al.* [128] use MAP as an important criterion for suggesting routes to users. Moreover, Dai *et al.* [37] and Campigotto *et al.* [83] discussed in our previous section can also be categorized

TABLE 5. Various parameters/factors proposed for personalized route recommendation.

Ref #	User Interest/ Rating	Time	Cost/Money/ Energy	Season	Distance	Visited Places	Type of Dataset	Dataset Website/Social Media Platform	Important Criterion for Routing
[37]		✓	✓		✓		Real	GPS/Trajectories	User Rating
[83]		✓	✓		✓		Real	Real People	Travel choices
[96]	✓	✓	✓	✓			Real	IgoUgo	MAP
[128]						✓	Real	Flickr API	MAP
[68]	✓				✓	✓	Real	Flickr	MAP
[123]	✓	✓				✓	Real	Panoramio	Directed graph with user preferences
[86]	✓						Real	Sina Weibo (aka Twitter of China)	Mean Absolute Error (MAE)
[11]	✓				✓	✓	Real and Non-Real	Gowalla	Matrix Factorization
[73]	✓					✓	Real	Foursquare and Whrrl	User preference + social and geographic influence
[33]	✓						Real	Foursquare	User Sentiment
[127]	✓						Real	Yelp	Highest User Rating
[87]	✓						Real	Flixster, MovieLens	Utility
[124]	✓						Real	Yelp, MovieLens, Douban Movie	MAE

under ''personalized route recommendation based on social datasets'' and thus are highlighted in Table 5.

4) BI-DIRECTIONAL ROUTE PLANNING USING TAXI GPS TRACES [10]

Here we discuss a very interesting case study-based approach presented in [10]. The reason we separate this approach from above presented strategies is because; (1) the authors present a very interesting case study for routing the night buses, and (2) the real-time data of various taxies is taken to plan the routes (which was missing in the strategies presented in above subsection). Moreover, buses are considered as one of the cheapest ways of traveling, therefore, providing automated routes to bus drivers as well as real-time information to passengers on their smartphones or similar devices would be a huge benefit to modern society.

By considering the above, Chen *et al.* [10] divide the night buses routing problem into two subproblems: *the candidate bus stop identification and the best bidirectional bus route selection.* To address the former, whole city is divided into equal sized small grids where each cell is 10 X 10 meters. Then, the cells with highest PDR (pick-up and drop-off rate) are combined to form a big cluster which is later divided into small sized clusters each having a principle bus station which is at a walking distance for a user within the specified cluster. A cluster merging algorithm is also presented allowing a big

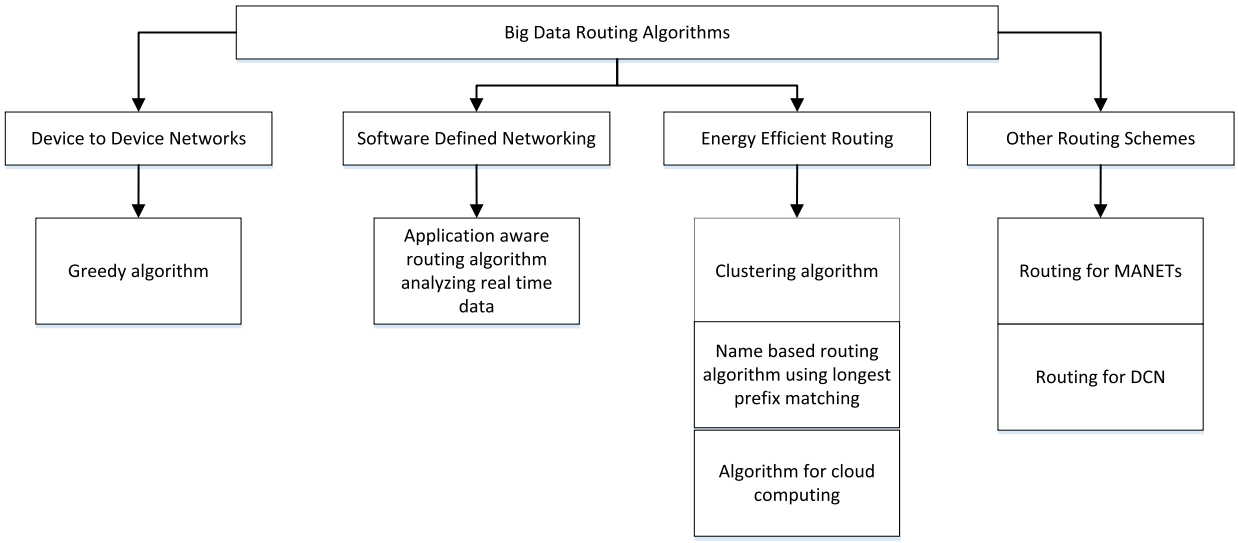

FIGURE 4. Division of existing big data routing algorithms in different categories.

cluster to merge nearby small clusters if their distance is not greater than a certain limit (i.e., if the absorbed cluster is not too far).

For bus route selection, the authors propose a graph theory based three-step process as follows:

- Build the graph with bus route and remove unnecessary nodes and edges. This process requires certain proposed criteria to be met such as the maximum distance between two stops is not too high (and is adequate), the bus should always move forward (farther) from the origin and close to destination, and the zigzag routes should be avoided ensuring the smoothness of the routes. Heuristics-based mathematical equations are also provided for each criterion.
- The information about the bus and its passengers is recorded in two different matrices known as Time and Flow matrices, respectively. Each element in time matrix represents the bus travel time from one stop to
- another. Likewise, each entry of Flow matrix corresponds to the number of passengers.
- Based on above two factors, the authors propose probability algorithm that generates best/optimal route for the passengers. The algorithm selects routes with large number of passengers and less distance. Similar sort of algorithm can also be found in [93].

All the theoretical aspects of [10] are supported by extensive set of experimental results conducted in MATLAB on real-time SBD considering thousands of taxis in Hangzhou, China over a period of one month. Interesting graphical illustrations are presented showing the convergence of proposed algorithm, optimal route selection based on less number of stops, direct relationship between number of stops, and the total traveling time for the buses. The results are compared to the algorithm proposed in [93] reflecting that the proposed algorithm converges to an optimal value.

B. ROUTING ALGORITHMS FOR MODERN DAY BIG DATA NETWORKS

Here we consider routing algorithms designed for BD traffic generated in modern day networks. We focus on various types of networks such as cognitive radio (CR) networks, D2D communications, SDN, energy efficient networks, MANETs (mobile ad hoc networks) and so on. The work cited in this section is not limited to the aforementioned types of networks and can easily be adapted to all kinds of traditional wireless networks. Our classification for this section is summarized in figure 4.

1) BIG DATA ROUTING IN COGNITIVE RADIO DEVICE TO DEVICE (CRD2D) NETWORKS [42]

D2D communication [70] plays an important role in BD networks due to its ad-hoc nature, flexible and light weight architectures, and anytime availability especially in critical conditions where the network resources are merely available. Huang *et al.* [42] address this interesting concept combining D2D communications with cognitive radio (CR) technology that exploits the unused licensed and unlicensed freely available spectrum [113]. This work falls within the scope of our survey since the devices movement is related to the movement of a person from one geographical location to another and thus, the routing algorithm is proposed based on real-time mobility of people. Basically, by integrating D2D and CR technologies, a ''socially aware'' big data (greedy) routing algorithm is proposed which uses the mobility of devices and spectrum as benefits for efficient data delivery. Here the mobility of nodes refers to the movement of people from one location to another which is finite and follows a periodic pattern. Same trend applies to the mobility of spectrum which is based on daily activities of licensed or primary users and thus can be predicted having finite patterns. In the proposed ''greedy'' algorithm, a node 'x' chooses the most

optimal node in the neighborhood which then acts as a ''relay'' to forward the data of 'x'. This optimal selection and relaying process continues unless the data reaches its final destination. Besides the proposed algorithm, the authors describe in detail the concept of CRD2D networks and identify open issues related to routing in these networks (such as effects of weak links on delivery of data, nodes communication overhead, and traffic control) which remained untouched in [41].

2) APPLICATION AWARE BIG DATA ROUTING IN SDN

SDN has shown great emergence in the recent past. Basically, in SDN, a centralized entity (a software defined centralized controller) or a platform controls the flow of all the network traffic [80]. The centralized controller adds flexibility in routing by intelligently controlling the flow of network traffic, thus making the integration of SDN with BD applications a feasible approach. The work proposed in [64] utilizes the ''application awareness'' advantage of SDN to perform real-time routing. Basically, the authors analyze the traffic patterns of MapReduce [69] based on the information provided by Hadoop (a well-known BD platform) [19]. An entity named as TaskTracker is allowed to store the information about traffic size and volume, which is later communicated to the SDN controller. This novelty was missing in [51] where the job of TaskTracker was only to send empty hello/heartbeat messages. Based on the information provided by Hadoop (via TaskTracker), the SDN controller performs routing of BD traffic by assigning priorities to one-on-one flows which are remotely located. Flows having all-to-all connectivity are given lower priorities since they have higher freedom in accessing the available bandwidth.

For experiments, real-world traffic patterns are generated using sixteen different Hadoop servers with a bandwidth of 10 Mbps and a delay of 1 ms, respectively. Most prominent feature of the results is the comparison of proposed SDN aware routing with existing solutions not utilizing the SDN as a part of their architecture (such as the famous spanning tree algorithm [129]). As expected, all results show the proposed approach significantly reduces the overall traffic shuffling time due to the flexibility added by the SDN controller. Nevertheless, there is no discussion nor any graphical illustration on other important parameters such as the system complexity, number of messages sent and received between Hadoop and SDN controller, and the effect of increased message size on global system performance.

3) ROUTING OF BIG DATA WITH ENERGY EFFICIENCY

Wireless sensor networks normally produce large amount of data by collecting information from various sensor nodes geographically located in a region or an area. This data is sometimes so large in size that it can easily be termed as big data. Din *et al.* [94] believe that this ''multi-sensor'' BD needs to be processed and routed by introducing a new ''cluster-based'' scheme that also keeps the battery usage level of sensory nodes at minimal. In the proposed approach (figure 5), sensor nodes gather (or sense) the information and forward it to the Base station (BS). This process is known as 'self-organization' where nodes learn about their neighbors through broadcasting. The responsibility of forming clustering group remains with the BS which is shown as 'flat layer design' in figure 5. The cluster head manages the communication with the BS, floods control messages to neighbors, selects cluster members, and chooses the most energy efficient path for routing as shown with steps 3 'cluster layer design' and 4 'cluster member selection'. Thus, all other sensor nodes are exempted from the overhead of optimal route selection. This in turn saves the battery life of senor nodes. Another novelty of the proposed approach is that only those devices form clusters which are at one-hop distance from the BS, thus the directly connected nodes to BS remain ''unclustered''. This also saves energy of sensor nodes. Once the clusters are formed, BD collected by the participating sensors has to be routed (or disseminated) to the desirable destinations. For this, a technique known as ''data fusion (step 5)''[1] is used, which separates data at various clusters both contextually and structurally. Fusion technique also monitors each cluster head's data processing performance. Results show that the proposed approach has high energy efficiency because; 1) the direct neighbors of BS remain ''unclustered'' which reduces the number of clusters, and 2) limited number of broadcast messages are exchanged only with selected neighbors. We also highlight the important steps of [94] in figure 5.

Another related approach with the focus on energy efficient routing for BD networks is presented in [89]. The entire focus of the proposed solution roams around utilizing the benefits of a content centric network (CCN) [29] for BD retrieval and sharing. An aggregatable named based routing (ANBR) algorithm is proposed that searches the closest copy of data for a user's request. Longest prefix matching is used to retrieve the desired copy of data requested by the user. Most of the routing information is stored locally which speeds up the retrieval process. Moreover, the authors claim that ANBR is energy efficient because it employs a hierarchical routing structure and uses longest prefix matching for information retrieval where one prefix can serve as a set of data to reduce the size of a routing table. This energy efficiency claim has been proven via simulation results, however, the authors failed to provide any comparison with any existing approach. Another promising energy efficient routing algorithm is reported in [109] with emphasis on BD cloud computing. The authors use extensive mathematical calculations with linear and dynamic approaches to model/calculate the power consumption of the network, network topology, and energy and time required for route computation. The proposed models have been evaluated using real-time statistics of four different routes each having distinct equipment, capacity, and power consumption. The results show that despite having large number of traversed nodes, one of the four routes consumes less

[1] We detail fusion in Section V.

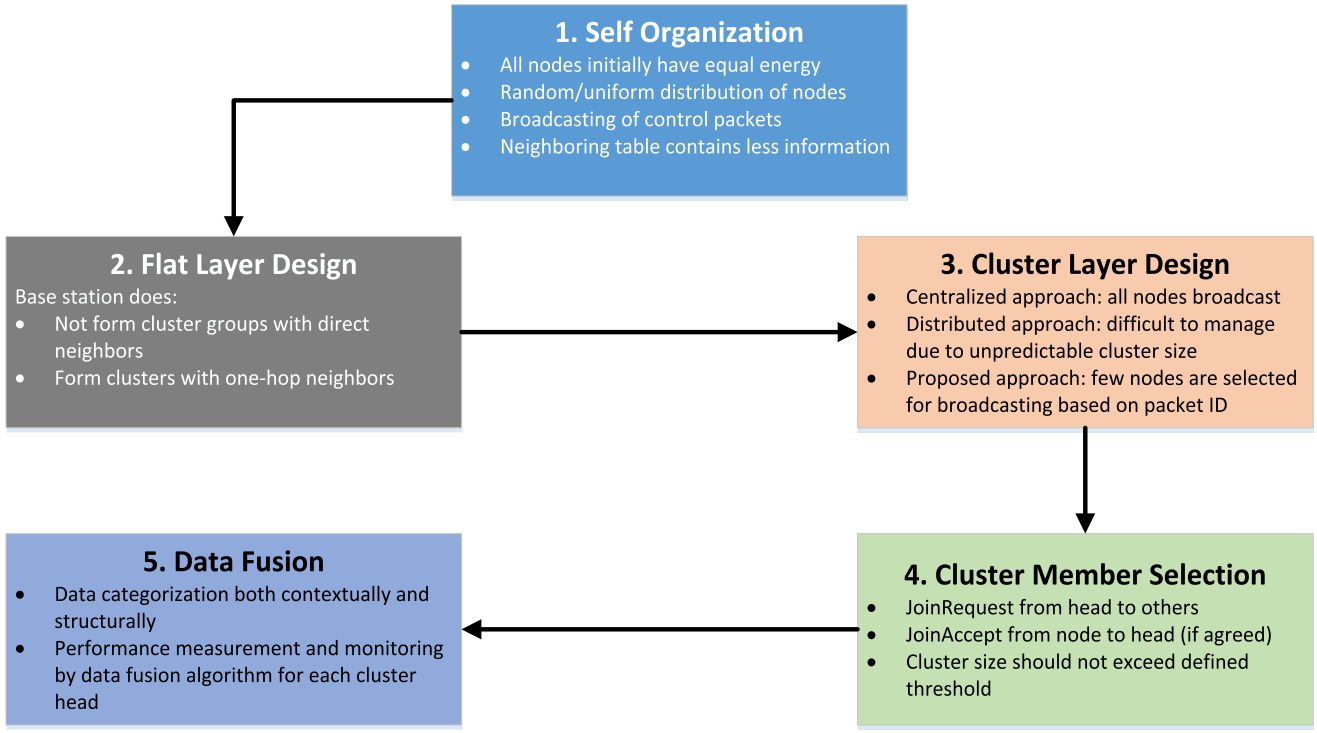

FIGURE 5. Five functional elements proposed in [94].

energy than the others. *This is due to the difference in capacity and power consumption of the equipment used*in each route.

4) OTHER IMPORTANT ROUTING ALGORITHMS FOR VARIOUS NETWORKS

Beside the algorithms presented so far, there are some important research efforts on routing of BD which are worth mentioning here. For example, the work presented in [4] considers analysis of BD traffic generated in MANETs using quality assurance metrics (such as functionality and usability of data). Big data carried by the traditional DSR (dynamic source routing) and AODV (ad hoc on demand distance vector) routing protocols has been simulated using NS2 (network simulator 2). Results depict that AODV performs better than DSR when combined with BD analytics. Moreover, another work presented in [26] examines the BD traffic in DCN (data center networks) [119]. The authors tackle a challenging task of determining the conditions in which the traditional routing algorithms can carry BD packets with less loss and high throughput. Thus, Markov chains are used for system analysis and probability estimations. Later, through simulations, it is shown that the time taken to reach a "non-blocking routing" situation is exponential. In addition, the proposed Markov model achieves a good convergence rate over a DCN of large number of nodes due to its simplicity in signaling. This signaling is named as EAR (explicit adaptation requests) in [26] and is performed between various switches to control the packet flow.

V. BIG DATA ANALYSIS FOR MOVING OBJECTS

There are different sources for data generation in moving objects including video devices, sensors, vehicles, etc. The obtained data serve as an input for real-time traffic analysis and decision making; therefore, it is not appropriate to rely only on these sources for data generation. Moreover, the recent advances in data analysis within the context of moving objects focus on proactive control and management [119]. For example, in vehicular communications, the analysis of real-time as well as historical data might result in accurate prediction of collisions and accidents. In the following sub-sections, we discuss several BD learning techniques for moving objects. In particular, we focus on learning techniques used in vehicular communications and intelligent transport system (ITS).

A. LEARNING DRIVEN TECHNIQUES

The existing learning-based mechanisms utilize different metrics which serve as an input for the analysis component [49]. Although, these metrics (such as trip duration) might be useful, however the challenge is that they depend on the traffic conditions, which are highly unpredictable and change over time. Therefore, these shallow traffic prediction models are not reliable for accurate modeling of the problem and real-world applications [72]. To solve the problem, van Lint [49] propose a new traffic prediction model which utilizes the temporal and spatial correlations. This learning model uses generic traffic flow features in a layered fashion. The area of concern for this model is high computational cost and limited scalability.

In [34], a predictor-corrector Kalman filter for travel and arrival time predictions is proposed. The authors use mathematical model with the node-arc network in which arcs represent the roadways while nodes represent the junctions. However, the focus of this contribution was more on mathematical modeling instead of actual traffic flows and condition in the network. Linda and Manic [81] use the active nearest neighbor for the identification of important places. In the second phase of the proposed algorithm, fuzzy logic is employed to compute the risk-level for the places identified in the first step.

For the purpose of autonomous navigation in cities, it is extremely important to analyze the trajectory pattern of the vehicles. One of the hurdles in such analysis is the inefficiency of the algorithms that rely on offline trajectory data [75]. This means after completing the learning stage, these algorithms are unable to learn diverse patterns. The solution to this problem is provided in [22] which focuses on incremental learning of trajectory patterns along with the predictions. The authors propose a growing hidden Markov model in which the constraints are derived through online learning for efficient identification of new trajectories. Angkititrakul *et al.* [82] predict the trajectory of the vehicle using Gaussian model by considering the parameters such as lane-crossing and driver correction events as inputs for learning.

Several automatic learning models are discussed in literature considering the mobility in vehicular and D2D networks [35], [122]. Veeraraghavan and Papanikolopoulos [35] propose a semi-supervised learning algorithm that represents the activities as sequence of actions and learns those activities as stochastic grammar. Similarly Mastronarde *et al.* [122] propose a supervised learning algorithm in which devices can learn the optimal cooperation strategy for providing the relaying services such as audio/video data. Another application of learning algorithms is routing decision in opportunistic networks [129] which uses decision trees and neural networks for efficient routing decisions. The parameters used for learning include routing scheme, device popularity, speed, location, and energy consumption.

B. DATA FUSION BASED TECHNIQUES

Another approach for achieving high performance in BD networks is to use multiple models. For example the work presented in [75] utilizes moving average, autoregressive moving average, and exponential smoothing model for traffic flow predictions. The predictions from these models are utilized in aggregation stage by the neural networks providing the ultimate prediction. Several algorithms fuse the predictions from different models for short-term flow prediction and trajectory changes [77], [92]. Toledo-Moreo and Zamora-Izquierdo [92] fuse the prediction from Kalman filtering in an equation of weighted coefficients. These weighted coefficients are generated in real-time during the prediction process. For efficient analysis and prediction of flow dynamics, Shawe *et al.* [48] suggest to utilize multi-source driven data fusion strategy. Such a strategy results in extremely broad and holistic information which increases the performance gain. Another data fusion approach is presented in [97] for surveillance and tracking of vehicles. The mechanism is based on robust data alignment that finds relational maps between the invariant feature datasets. These datasets consist of aligned images that are gathered from different unnamed aerial vehicles.

Several solutions focus on utilizing multi-source information fusion that combines core evidence for highly dynamic, diverse, and conflicting data received through different sources [82], [88]. The ultimate objective is to extract a comprehensive estimate about an event by integrating information received from multiple sources. Angkititrakul *et al.* [82] analyze several approaches for fusing video streams and propose an algorithm for vehicle trajectory prediction considering the lane-crossing and driver correction events. One of the problems with the proposed approach is that it considers only the most common driving signals which limit the scope and application for adaptability. The approach used by Kong *et al.* [88] is based on multi-source information that includes fusing data from underground loop detector and GPS. The first part of the algorithm creates a credible and robust platform for the fusion of multi-source data, while, in the second part, wave theory is applied for estimating the mean speed. One of the issues with data collection using roadside sensors and GPS is that the collected data is sparse due to random distribution of vehicles in time and space. Mehta and Chana [117] discuss several techniques for flow estimation by considering parameters such as running time and accuracy of used data. It is observed that most of the data comes from probing (roadside sensors, GPS, etc.) at different sampling rates. The authors propose higher sampling interval which results in lower quantity and accuracy of data but provides efficiency in terms of processing.

Polychronopoulos *et al.* [6] propose a hierarchical structure algorithm for predicting the trajectory of moving objects using the fusion of environmental and vehicle dynamic data. García *et al.* [44] present a multi-sensory system for the detection of obstacles by using a set of diverse techniques for data fusion. Sun and Zhang [104] suggest a new approach which integrates the spatial and temporal information for the flow forecast. For selecting the input variables, they use Pearson correlation coefficient. The resulting multiple outputs are then integrated using a fusion algorithm for final predictions.

C. RULE EXTRACTION BASED TECHNIQUES

One of the important features for learning-driven moving object models is to understand the pattern, association, and correlation among various datasets acquired through different sources. The objective is typically achieved using association rule mining (ARM). Lin *et al.* [120] utilize the association analysis with ARM for the prediction of network flow. Similarly, ARM is used for the classification of network flow anomalies in [65]. Another technique for rule extraction is based on rough set theory (RST) [55], [56], [84].

RST extracts useful data from decision table using induction based decision-making techniques. BD is analyzed using statistical methods and association rule in order to identify the common attributes that appear with high frequency. One of the advantages of using RST is to extract important attributes without having any information outside the dataset.

D. ADP (ADAPTIVE DISTRIBUTED PROGRAMMING)-BASED LEARNING CONTROL

Although, learning based techniques seem effective and the results are also quite promising, yet realizing the performance in a highly dynamic and uncertain environment is quite difficult. Several mathematical approaches are proposed to handle such problems [27]. In order to solve complex decision processes having huge and continuous states, Wang *et al.* [27] propose methods that use reinforcement learning (RL) with dynamic programming. The dynamic programming algorithms are classified into two classes. First class is the one that does not need any initial stable policy while the other class of algorithms requires an initial stable policy. Ling and Shalaby [61] study various elements of RL agents such as reward, action set, and state space in the context of traffic flow. One of the most promising features of RL is the ability of an agent to understand the relationship between controlled actions and their consequences on the environment. Abdulhai *et al.* [8] utilize this feature and propose a Q-learning method for network traffic control. Salkham *et al.* [7] utilize the information provided by GPS and V2V (vehicle to vehicle) communication for traffic optimization. They use a local round robin switching model based on RL. A lot of work still needs to be done but it is expected that dynamic programming variants, methods, and algorithms will provide the basic functionality for realizing an optimized and dynamic system for moving objects.

E. OBJECT-BASED LEARNING

The object movement in V2V, sensor networks, or D2D networks has its own unique characteristic which is extremely important for learning driven models and algorithms. The spatial temporal relationship in traffic flow data and geographical information are important considerations for efficient learning-driven systems. It is important to distinguish between two close but unrelated data points that may wrongly be clustered in the same group [103], [126]. Similarly, the persistent congestion in V2V networks might arise due to many reasons including the network congestion, accidents, and so forth [3]. The same scenarios are true for sensor networks as well as the other networks involving moving objects. Therefore, it is extremely important to incorporate diverse patterns in the learning systems which are ultimately applicable to real-life scenarios.

F. BIG DATA ANALYSIS FOR SOCIAL MEDIA

Social media and data analysis techniques are extremely important for effective proliferation of user-generated content. The collection of data from social media and its analysis using existing BD tools help in mining users sentiment which results in effective marketing and service activities. The large amount of data produced by the social media users cannot be handled by conventional data management techniques. There is a need of real-time analytics and powerful metrics for the analysis of social media data. The efficient analysis of social media data has created new opportunities to understand and influence how people think and act.

The huge volume of social media data combined with the mobility of objects needs fast processing and analysis. The existing data analysis techniques can handle volume, velocity, and variety of data and mining of significant patterns from the data. In this regard, Godbole *et al.* [172] discuss the semantic orientation of words and sentences on social media. Adjectives divided with AND are known to have the equal polarity while the ones divided by BUT have reverse polarity. Adjectives of the same affinity can be grouped into clusters using statistical model. Another major problem is the summarizing of huge volume of data generated by social media. Ku *et al.* [173] proposed opinion summarization by analyzing the sentiment polarities, degree, and the associated occurrences. However, one of the drawbacks for such approach is that not all opinions are of same importance. For example, the nature of news is quite different from the blog articles. Compared with blog articles, news articles own a larger vocabulary making it harder to retrieve the relevant sentence.

In addition to above, several clustering techniques are used for analyzing social media information of moving objects [174], [175]. Kaur and Singh [175] presented a wide variety of approaches for anomaly detection in social network such as structured, clustering, and Bayesian classifier based techniques. However, despite having this wide variety of work there are a few shortcomings. First, temporal constraints (such as previous information for learning the defined model) need to be added for analyzing social media. Secondly, the analysis of big data in social networks for the presence of anomalies requires a lot of attention. It is important to focus on unstructured behavior of data in social networks rather than predefined set of labeled data or randomly chosen nodes.

Table 6 provides the description, issues, and challenges for the BD analysis techniques discussed in sections V and VI.

In the next section, we analyze some important research issues that are essential for an efficient data driven moving object system.

VI. ISSUES AND FUTURE DIRECTIONS FOR BIG DATA-DRIVEN MOVING OBJECTS

In the context of V2V communication and BD routing, the use of technology such as GPS, sensors, and other IoT devices provides the important data for analysis. However, using a large number of these devices not only incurs huge cost but also results in redundant data. Another interesting aspect is related to the application of BD analysis on social media. Social media provides diverse information including audio, videos, messages on the fly, and other information generated by sensors attached with the mobile devices. Moreover, social

TABLE 6. Brief summary of data analysis techniques discussed in section V and VI.

Data Analysis Technique	Description	Reference	Issues	Solutions
Learning Driven	These techniques are based on the intrinsic mechanism for the moving objects using both real-time and historical data collected through different sources	[22][34][35] [49][72][75] [81][82] [122][129]	• Data cleansing on the actual received data is challenging. Removing of useless noisy parts is required • Multi-dimensional data increases the problem complexity of the analytics • Redundant features of data result in possible duplication • Heterogeneous data collected from wide variety of sources needs to be incorporated for specific scenarios	• Refining the data and removing the abnormalities [63][123]. • Manifold learning and dimension reduction [36][106] • Sparsity penalization and reduction in coding complexity [39][43] • Using procrustes analysis and canonical correlation [16][59]
Data Fusion	These techniques mostly use multiple models on BD for achieving high performance. The prediction from these models are utilized in aggregation stage	[6][44][75] [77][82][88] [92][97] [104][117]	• Energy consumption is one of the main concerns especially for small devices such as sensors • Latency and delay must be within considerable range for fast processing of data • Designed algorithms must be scalable having low time and space complexity	• Use of multi-source driven data fusion strategy providing holistic information • Multi-sensory system integrating spatial and temporal information for the flow forecast
Rule Extraction	Techniques to understand the pattern, association, and correlation among various datasets acquired through different sources	[55][56][65] [84][120]	• Realizing the performance in a highly dynamic and uncertain environment is quite difficult	• BD is analyzed using statistical methods and association rule in order to identify the common attributes that appear with high frequency [84]
ADP Based Learning Control	This technique uses ADP and reinforcement learning for the performance optimization of complex dynamic systems	[6][7][8] [27]	• Latency and delay must be within considerable range for fast processing of data.	• A lot of work still needs to be done but it is expected that dynamic programming variants, methods, and algorithms will provide the basic functionality for realizing an optimized and dynamic system for moving objects
Object Based Learning	The spatial-temporal relationship in traffic flow data and geographical information are important considerations for object based learning	[103][126]	• The persistent congestion in sensor and V2V networks might arise due to many reasons including the network congestion, accidents, and so forth [3]	• Incorporate diverse patterns such as vehicle type in the learning systems, which are ultimately applicable to real-life scenarios

media applications are one of the major sources for BD analysis but there are certain inherited challenges associated with this type of BD. Firstly, it is important to get huge number of users providing reliable data. At the same time, it is required to address the issues related to the security and privacy of users' data [132], [167]. The anonymity of the data must be

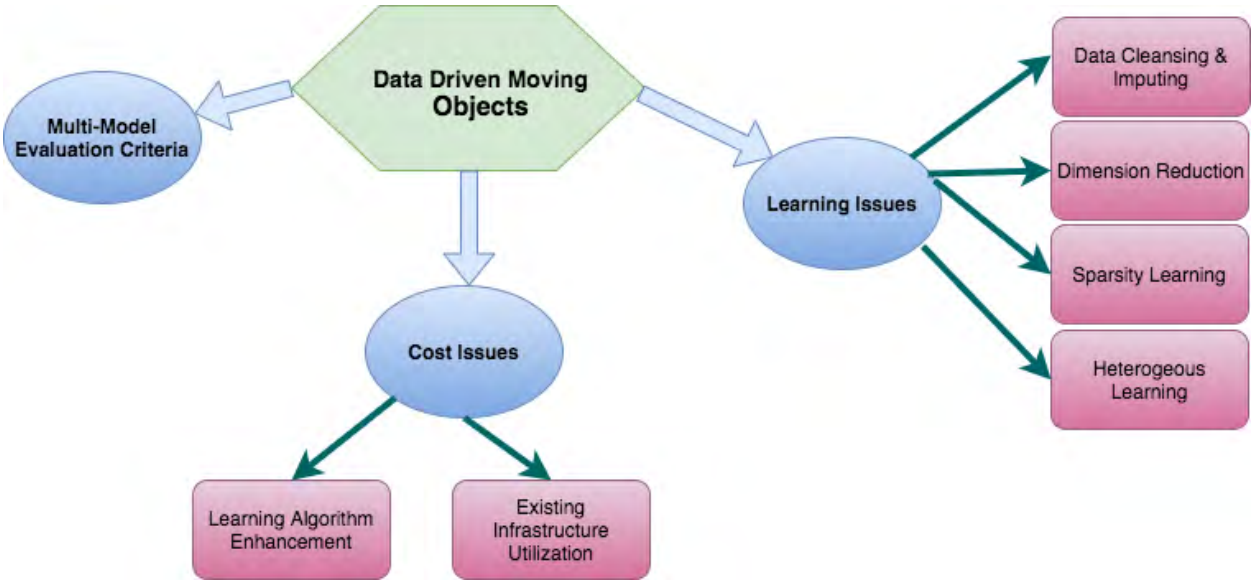

FIGURE 6. Issues for BD-Driven moving objects.

protected in such a way that not only the data makes sense but it also preserves the identity of the subjects [165], [166]. At some point of the time, the analysis of data provides useful information such as traffic routes and congestion notification. The model is a dynamic people-centric model.

As discussed in previous sections, data is the main element in any analysis framework. The accuracy and efficiency of data collected from different sources are extremely crucial for the analysis phase as well as the design of routing algorithms. While using clustering and data fusion techniques, the energy consumption is an important concern for the sensory devices [94], [109]. Apart from energy consumption, latency and delay must be in considerable range for fast processing of data. Thus, the designed algorithms must be scalable and should have low time and space complexity [89]. In order to reduce the network contention, dynamic routing algorithms must be deployed instead of static ones [26].

In the following subsections, we will discuss some important problems related to the utilization of learning-based systems for moving objects and identify areas for future research as shown in figure 6. The figure shows the learning and cost issues which are explained in detail in the next sub-sections.

A. LEARNING ISSUES

The data gathered from different sources is mostly multi-dimensional, heterogeneous, and irregular [98]. This leads us to the following four challenges:

1) DATA CLEANSING AND IMPUTING

Data provided by different sources is full of noise [52], therefore, it is extremely important to perform the cleansing of data for removing the useless and noisy parts. The process of data cleansing is not straight forward. Lu *et al.* [123] propose a data analysis technique that also performs data cleansing by refining the data and removing the abnormal parts. The noise removal process here is based on statistical estimation information of several noise types. One of the drawbacks with these kinds of approaches is that the noise in reality is arbitrary while these approaches consider it as some known form. Probabilistic principal component analysis (PPCA) [63] is another method used to capture different features of traffic flow. It is also used to filter out the useless and abnormal data that effects the imputation process. PPCA based methods not only consider the local information but also rely on global information such as relationship between historical data.

2) DIMENSION REDUCTION

The complexity of learning problem is another issue in data analysis for moving objects. In case of V2V, the vehicle image can be considered having multiple dimensions. The increase in dimensions results in exponential increase in number of samples. There are several dimension reduction techniques proposed in [36] and [106]. An important method in this context is known as manifold learning [54]. The basic idea behind manifold is discovering the low dimensional manifold integrated in high dimensional Euclidean space. The initial inspiration of manifold is to obtain the nonnegative part from the data [12]. The other method for dimension reduction is Kernel dimension [58]. It uses supervised information to make decision on the dimension reduction. Such methods help to extract very useful information for moving object analysis and help to improve the performance of learning-driven tasks.

3) SPARSITY LEARNING

Sparse learning is different from dimension reduction in a sense that it focuses on the removal of certain redundant features (such as dimension reduction of interference data) while preserving the originality of the remaining

feature [91], [111]. Tibshirani [91] identifies these redundant features and give them extremely low weightage for ensuring smooth removal. Several refinements to this procedure are proposed later in [32] and [74]. Duchi and Singer [39] utilize sparsity penalization into the boosting algorithms for attaining higher performance. Although, there are several methods for employing the coding complexity related to the structure of the featured set [43], it is still an important concern to differentiate redundant features and the features crucial for the performance.

4) HETEROGENEOUS LEARNING

Another important issue is the heterogeneous property of data that is collected from a wide variety of sources. Data fusion of such heterogeneous data is an extremely challenging task. Although there are several ways to handle the heterogeneous data but two common methods are based on machine learning. The first method uses common space such as Procrustes analysis and canonical correlation [16], [59]. These methods assume the linearity of the datasets transformation. Kernel stick is used as another method for transforming linear data to non-linear [95].

B. COST ISSUES

Cost is still a major concern for data analysis in moving objects. One of the major concerns in this regard is the data collection sources. It is possible to reduce the cost in BD routing by designing an efficient classifier for object recognition [38], [53]. Another solution is to utilize low cost alternative devices while improving the performance of the algorithms. It is also possible to use the existing infrastructure deployed for other applications [53], [62]. Clanton *et al.* [45] combine the high accuracy map and the navigation system to propose a lane departure system. Such system reduces the cost by eliminating the requirement for expensive GPS receivers. Similarly, the probe cell phones are used to estimate the travel time between two points. The advantage here is to use the already deployed infrastructure for acquiring the desired data [62].

C. MULTIMODAL EVALUATION CRITERIA

It is important to measure the effectiveness of any method using different evaluation criteria. The data in moving objects is highly dynamic and varies over time so a static metric may not provide the complete picture [60]. There are different algorithms that consider multimodal evaluation criteria and argue that a single or a certain group of criteria might not be effective for all cases [71]. Hussein *et al.* [71] evaluated different algorithms using detection error tradeoff curves. It was observed that the detection performance is not influenced due to different sensors however; the window size chosen for modeling the classifiers has significant impact on detection performance. In order to identify the problem in detail and propose a model for solution, we are required to establish a multimodal evaluation criterion considering the dynamic nature of data sources as well as the moving objects [2]. Multimodal evaluation criterion is not important for identifying the problem however, it helps in effective design of the system.

VII. CONCLUSION

Recent years have shown a tremendous growth in volume and complexity of spatially generated big data. This growth is due to the decline in prices of portable devices as well as the 24/7 availability of internet connections. The handling and processing of such immensely generated data become more complex when the movement of spatial objects is random and dynamic. These objects may range from humans to moving vehicles, geographically located sensors to objects and moving animals with RFID chips, and so forth. The data generated by these objects is not small in size, for instance, a person traveling in a car might use several applications together which need real-time analysis and processing. The amount and complexity of this data double with inclusion of modern day social media applications. In this paper, we tried to highlight various important approaches, algorithms, and trends related spatial big data (SBD) and moving objects. We have surveyed sources of SBD generation and highlighted their routing, analysis, and processing techniques. We have observed that there is a variety of techniques to deal with complexity of SBD ranging from personalized routing to data fusion, however still a lot of research has to be done in future for more advanced real-time applications. We concluded our survey by outlining open issues and challenges with some future directions for big data driven moving objects.

REFERENCES

[1] A. B. T. Sherif, K. Rabieh, M. M. E. A. Mahmoud, and X. Liang, "Privacy-preserving ride sharing scheme for autonomous vehicles in big data era," *IEEE Internet Things J.*, vol. 4, no. 2, pp. 611–618, Apr. 2016.

[2] A. Hobeika and N. Yaungyai, "Evaluation update of the red light camera program in Fairfax County, VA," *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 4, pp. 588–596, Dec. 2006.

[3] A. K. H. Tung, J. Hou, and J. Han, "Spatial clustering in the presence of obstacles," in *Proc. Int. Conf. Data Eng.*, 2001, pp. 359–367.

[4] A. Kush, C. J. Hwang, and V. Dattana, "Big data analytics on MANET routing standardization using quality assurance metrics," in *Proc. IEEE Future Technol. Conf. (FTC)*, 2016, pp. 192–198.

[5] A. Masini, G. Corsini, M. Diani, and M. Cavallini, "Analysis of multiresolution-based fusion strategies for a dual infrared system," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 4, pp. 688–694, Dec. 2009.

[6] A. Polychronopoulos, M. Tsogas, A. J. Amditis, and L. Andreone, "Sensor fusion for predicting vehicles' path for collision avoidance systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 3, pp. 549–562, Sep. 2007.

[7] A. Salkham, R. Cunningham, A. Garg, and V. Cahill, "A collaborative reinforcement learning approach to urban traffic control optimization," in *Proc. Web Intell. Intell. Agent Technol.*, 2008, pp. 560–566.

[8] B. Abdulhai, R. Pringle, and G. J. Karakoulas, "Reinforcement learning for *True* adaptive traffic signal control," *J. Transp. Eng.*, vol. 129, pp. 278–285, May/Jun. 2003.

[9] Big Data Definition. Gartner, Inc. [Online]. Available: http://www.gartner.com/it-glossary/big-data/

[10] C. Chen, D. Zhang, N. Li, Z.-H. Zhou, "B-planner: Planning bidirectional night bus routes using large-scale taxi GPS traces," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 4, pp. 1451–1465, Aug. 2014.

[11] C. Cheng, H. Yang, I. King, and M. R. Lyu, "Fused matrix factorization with geographical and social influence in location-based social networks," in *Proc. 26th AAAI Conf. Artif. Intell.*, 2012, pp. 17–23.

[12] C. H. Q. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 45–55, Jan. 2010.

[13] C. Jardak, P. Mähönen, and J. Riihijärvi, "Spatial big data and wireless networks: Experiences, applications, and research challenges," *IEEE Netw.*, vol. 28, no. 4, pp. 26–31, Jul./Aug. 2014.
[14] C. S. Jensen, "Data-intensive routing in spatial networks," in *Proc. ACM TURC*, 2017.
[15] C. Snijders, U. Matzat, and U.-D. Reips, "'Big data': Big gaps of knowledge in the field of Internet science," *Int. J. Internet Sci.*, vol. 7, no. 1, pp. 1–5, 2012.
[16] C. Wang and S. Mahadevan, "Manifold alignment using procrustes analysis," in *Proc. Int. Conf. Mach. Learn.*, 2008, pp. 1120–1127.
[17] C. Liu, S. Zhang, H. Wu, and Q. Fu, "A dynamic spatiotemporal analysis model for traffic incident influence prediction on urban road networks," *Int. J. Geo-Inf.*, vol. 6, no. 11, p. 362, 2017.
[18] C. Wu, A. Kreidieh, K. Parvate, E. Vinitsky, and A. M. Bayen. (Oct. 2017). "Flow: Architecture and benchmarking for reinforcement learning in traffic control." [Online]. Available: https://arxiv.org/abs/1710.05465
[19] D. Cheng, X. Zhou, P. Lama, M. Ji, and C. Jiang, "Energy efficiency aware task assignment with DVFS in heterogeneous hadoop clusters," *IEEE Trans. Parallel Distrib. Syst.*, vol. 29, no. 1, pp. 70–82, Jan. 2018.
[20] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, Oct. 1999.
[21] D. Laney, "3D data management: Controlling data volume, velocity and variety," META Group, Inc, Tech. Rep., Feb. 2001. [Online]. Available: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf
[22] D. Vasquez, T. Fraichard, and C. Laugier, "Incremental learning of statistical motion patterns with growing hidden Markov models," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 3, pp. 403–416, Sep. 2009.
[23] D. Xia, B. Wang, H. Li, Y. Li, and Z. Zhang, "A distributed spatial–temporal weighted model on MapReduce for short-term traffic flow forecasting," *Neurocomputing*, vol. 179, pp. 246–263, Feb. 2016.
[24] *Draft NIST Big Data Interoperability Framework: Volume 1, Definitions*, NIST, Gaithersburg, MD, USA, 2015.
[25] E. Chaniotakis and C. Antoniou, "Use of geotagged social media in urban settings: Empirical evidence on its potential from Twitter," in *Proc. IEEE 18th Int. Conf. Intell. Transp. Syst.*, Sep. 2015, pp. 214–219.
[26] E. Zahavi, I. Keslassy, and A. Kolodny, "Distributed adaptive routing for big-data applications running on data center networks," in *Proc. ACM/IEEE Symp. Archit. Netw. Commun. Syst. (ANCS)*, Oct. 2012, pp. 99–110.
[27] F.-Y. Wang, H. Zhang, and D. Liu, "Adaptive dynamic programming: An introduction," *IEEE Comput. Intell. Mag.*, vol. 4, no. 2, pp. 39–47, May 2009.
[28] F. Xia, W. Wang, T. M. Bekele, and H. Liu, "Big scholarly data: A survey," *IEEE Trans. Big Data*, vol. 3, no. 1, pp. 18–35, Mar. 2017.
[29] G. Edens *et al.*, "A better way to organize the Internet: Content-centric networking," in *Proc. Blog IEEE Spectr.*, 2017.
[30] G. Pauer, "Development potentials and strategic objectives of intelligent transport systems improving road safety," *Transport Telecommun. J.*, vol. 18, no. 1, pp. 15–24, 2017.
[31] G. Zhao, X. Qian, and X. Xie, "User-service rating prediction by exploring social users' rating behaviors," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 496–506, Mar. 2016.
[32] G.-J. Qi, J. Tang, Z.-J. Zha, T.-S. Chua, and H.-J. Zhang, "An efficient sparse metric learning in high-dimensional space via ℓ_1-penalized log-determinant regularization," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 841–848.
[33] H. Gao, J. Tang, X. Hu, and H. Liu, "Content-aware point of interest recommendation on location-based social networks," in *Proc. 29th Int. Conf. AAAI*, 2015, pp. 1–7.
[34] H. Jula, M. Dessouky, and P. A. Ioannou, "Real-time estimation of travel times along the arcs and arrival times at the nodes of dynamic stochastic networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 9, no. 1, pp. 97–110, Mar. 2008.
[35] H. Veeraraghavan and N. P. Papanikolopoulos, "Learning to recognize video-based spatiotemporal events," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 4, pp. 628–638, Dec. 2009.
[36] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000.
[37] J. Dai, B. Yang, C. Guo, and Z. Ding, "Personalized route recommendation using big trajectory data," in *Proc. IEEE ICDE*, pp. 543–554, 2015.
[38] J. Du and M. J. Barth, "Next-generation automated vehicle location systems: Positioning at the lane level," *IEEE Trans. Intell. Transp. Syst.*, vol. 9, no. 1, pp. 48–57, Mar. 2008.
[39] J. Duchi and Y. Singer, "Boosting with structural sparsity," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 297–304.
[40] J. Gantz and D. Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east," IDC, Framingham, MA, USA, Study Rep., Dec. 2012. [Online]. Available: http://www.emc.com/leadership/digital-universe/index.htm
[41] J. Huang, S. Wang, X. Cheng, M. Liu, Z. Li, and B. Chen, "Mobility-assisted routing in intermittently connected mobile cognitive radio networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 11, pp. 2956–2968, Nov. 2013.
[42] J. Huang, S. Wang, X. Cheng, and J. Bi, "Big data routing in D2D communications with cognitive radio capability," *IEEE Wireless Commun.*, vol. 23, no. 4, pp. 45–51, Aug. 2016.
[43] J. Huang, T. Zhang, and D. Metaxas, "Learning with structured sparsity," in *Proc. Int. Conf. Mach. Learn.*, vol. 12, pp. 417–424, Jan. 2009.
[44] J. J. García *et al.*, "Efficient multisensory barrier for obstacle detection on railways," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 3, pp. 702–713, Sep. 2010.
[45] J. M. Clanton, D. M. Bevly, and A. S. Hodel, "A low-cost solution for an integrated multisensor lane departure warning system," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 1, pp. 47–59, Mar. 2009.
[46] J. Manyika *et al.*, "Big data: The next frontier for innovation, competition, and productivity," McKinsey, New York, NY, USA, Tech. Rep., May 2011.
[47] J. S. Ward and A. Barker. (Mar. 2014). "Undefined by data: A survey of big data definitions." [Online]. Available: https://arxiv.org/abs/1309.5821
[48] J. Shawe-Taylor, T. D. Bie, and N. Cristianini, "Data mining, data fusion and information management," *IEE Proc. - Intell. Transport Syst.*, vol. 153, no. 3, pp. 221–229, Sep. 2006.
[49] J. W. C. Van Lint, "Online learning solutions for freeway travel time prediction," *IEEE Transp. Intell. Transp. Syst.*, vol. 9, no. 1, pp. 38–47, Jan. 2008.
[50] J. Wang, Y. Wu, N. Yen, and S. Guo, and Z. Cheng, "Big data analytics for emergency communication networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1758–1778, 3rd Quart., 2016.
[51] J. Yu, F. Jiang, and T. Zhu, "RTIC-C: A big data system for massive traffic information mining," in *Proc. Cloud Comput. Big Data (CloudCom-Asia)*, Dec. 2013, pp. 395–402.
[52] J. Zhang, D. Chen, and U. Kruger, "Adaptive constraint k-segment principal curves for intelligent transportation system," *IEEE Trans. Intell. Transp. Syst.*, vol. 9, no. 4, pp. 666–677, Dec. 2008.
[53] J. Zhang, J. Pu, C. Chen, and R. Fleischer, "Low-resolution gait recognition," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 4, pp. 986–996, Aug. 2010.
[54] J. Zhang, H. Huang, and J. Wang, "Manifold learning for visualizing and analyzing high-dimensional data," *IEEE Intell. Syst.*, vol. 25, no. 4, pp. 54–61, Jul./Aug. 2010.
[55] J.-R. Chang, C.-T. Hung, G.-H. Tzeng, and S.-C. Kang, "Using rough set theory to induce pavement maintenance and rehabilitation strategy," in *Rough Sets and Knowledge Technology* (Lecture Notes in Computer Science), vol. 4481. Berlin, Germany: Springer-Verlag, 2007, pp. 542–549.
[56] J.-T. Wong and Y.-S. Chung, "Rough set approach for accident chains exploration," *Accident Anal. Prevention*, vol. 39, no. 3, pp. 629–637, 2007.
[57] K. Boriboonsomsin, M. J. Barth, W. Zhu, and A. Vu, "Eco-routing navigation system based on multisource historical and real-time traffic information," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 4, pp. 1694–1704, Dec. 2012.
[58] K. Fukumizu, F. R. Bach, and M. I. Jordan, "Kernel dimension reduction in regression," *Ann. Statist.*, vol. 37, no. 4, pp. 1871–1905, 2009.
[59] K. Fukumizu, F. R. Bach, and A. Gretton, "Statistical consistency of kernel canonical correlation analysis," *J. Mach. Learn. Res.*, vol. 8, pp. 361–383, Feb. 2007.
[60] K. G. Zografos and K. N. Androutsopoulos, "Algorithms for itinerary planning in multimodal transportation networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 9, no. 1, pp. 175–184, Mar. 2008.
[61] K. Ling and A. S. Shalaby, "A reinforcement learning approach to streetcar bunching control," *J. Intell. Transp. Syst., Technol., Planning, Oper.*, vol. 9, no. 2, pp. 59–68, 2005.

[62] K. Sohn and K. Hwang, "Space-based passing time estimation on a freeway using cell phones as traffic probes," *IEEE Trans. Intell. Transp. Syst.*, vol. 9, no. 3, pp. 559–568, Sep. 2008.

[63] L. Qu, L. Li, Y. Zhang, and J. Hu, "PPCA-based missing data imputation for traffic flow volume: A systematical approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, pp. 512–522, Sep. 2009.

[64] L.-W. Cheng and S.-Y. Wang, "Application-aware SDN routing for big data networking," in *Proc. IEEE GLOBECOM*, Dec. 2015, pp. 1–6.

[65] L. Trajkovic *et al.*, "Data mining and machine learning for analysis of network traffic," in *Proc. IEEE ICSSE*, Jul. 2017.

[66] M. Cox and D. Ellsworth, "Application-controlled demand paging for out-of-core visualization," in *Proc. IEEE Vis.*, Oct. 1997, pp. 235–244.

[67] M. Bakillah, S. H. L. Liang, A. Mobasheri, and A. Zipf, "Towards an efficient routing Web processing service through capturing real-time road conditions from big data," in *Proc. Comput. Sci. Electron. Eng. Conf. (CEEC)*, Sep. 2013, pp. 152–155.

[68] M. Clements, P. Serdyukov, A. P. de Vries, and M. J. T. Reinders. (2011). "Personalised travel recommendation based on location co-occurrence." [Online]. Available: https://arxiv.org/abs/1106.5213

[69] M. Goudarzi, "Heterogeneous architectures for big data batch processing in MapReduce paradigm," *IEEE Trans. Big Data*, to be published.

[70] M. Haus, M. Waqas, A. Y. Ding, Y. Li, S. Tarkoma, and J. Ott, "Security and privacy in device-to-device (D2D) communication: A review," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 1054–1079, 2nd Quart., 2017.

[71] M. Hussein, F. Porikli, and L. Davis, "A comprehensive evaluation framework and a comparative study for human detectors," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 3, pp. 417–427, Sep. 2009.

[72] M. Kubička *et al.*, "Performance of current eco-routing methods," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2016, pp. 472–477.

[73] M. Ye, P. Yin, W.-C. Lee, and D.-L. Lee, "Exploiting geographical influence for collaborative point-of-interest recommendation," in *Proc. ACM SIGIR*, 2011, pp. 325–334.

[74] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc., B (Statist. Methodol.)*, vol. 68, no. 1, pp. 49–67, 2006.

[75] M.-C. Tan, S. C. Wong, J.-M. Xu, Z.-R. Guan, and P. Zhang, "An aggregation approach to short-term traffic flow prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 1, pp. 60–69, Mar. 2009.

[76] M. Gmira, M. Gendreau, A. Lodi, and J.-Y. Potvin, "Travel speed prediction using machine learning techniques," in *Proc ITS World Congr.*, Montreal, QC, Canada, Oct./Nov. 2017, pp. 1–10.

[77] M. Ma, S. Liang, H. Guo, and J. Yang, "Short-term traffic flow prediction using a self-adaptive two-dimensional forecasting method," *Adv. Mech. Eng.*, vol. 9, no. 8, 2017.

[78] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-preserving multi-keyword ranked search over encrypted cloud data," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 1, pp. 222–233, Jan. 2014.

[79] N. Khan *et al.*, "Big data: Survey, technologies, opportunities, and challenges," *Sci. World J.*, vol. 2014, Jul. 2014, Art. no. 712826.

[80] N. Kitsuwan, S. Ba, E. Oki, T. Kurimoto, and S. Urushidani, "Flows reduction scheme using two MPLS tags in software-defined network," *IEEE Access*, vol. 5, pp. 14626–14637, 2017.

[81] O. Linda and M. Manic, "Online spatio-temporal risk assessment for intelligent transportation systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, pp. 194–200, Mar. 2011.

[82] P. Angkititrakul, R. Terashima, and T. Wakita, "On the use of stochastic driver behavior model in lane departure warning," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, pp. 174–183, Mar. 2011.

[83] P. Campigotto, C. Rudloff, M. Leodolter, and D. Bauer, "Personalized and situation-aware multimodal route recommendations: The FAVOUR algorithm," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 1, pp. 92–102, Jan. 2017.

[84] P. Haluzová, "Effective data mining for a transportation information system," *Acta Polytech.*, vol. 48, no. 1, pp. 24–29, 2008.

[85] P. Illapani, "Big data for enterprise customers leveraging hadoop distributed file system and SAP real time data platform," SAP Hana Blog, Tech. Rep., 2012.

[86] P. Lou, G. Zhao, X. Qian, H. Wang, and X. Hou, "Schedule a rich sentimental travel via sentimental POI mining and recommendation," in *Proc. IEEE Multimedia Big Data (BigMM)*, Apr. 2016, pp. 33–40.

[87] Q. Yuan, G. Cong, and A. Sun, "Graph-based point-of-interest recommendation with geographical and temporal influences," in *Proc. ACM Int. Conf. Conf. Inf. Knowl. Manage. (CIKM)*, 2014, pp. 659–668.

[88] K. Qing-Jie *et al.*, "An approach to urban traffic state estimation by fusing multisource information," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 3, pp. 499–511, Mar. 2009.

[89] R. Li, H. Harai, and H. Asaeda, "An aggregatable name-based routing for energy-efficient data sharing in big data era," *IEEE Access*, vol. 3, pp. 955–966, 2015.

[90] R. Michael *et al.*, "Enabling spatial big data via CyberGIS: Challenges and opportunities," in *CyberGIS: Fostering a New Wave of Geospatial Innovation and Discovery*, S. Wang and M. Goodchild, Eds. 2014.

[91] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc., B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[92] R. Toledo-Moreo and M. A. Zamora-Izquierdo, "IMM-based lane-change prediction in highways with low-cost GPS/INS," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 1, pp. 180–185, Mar. 2009.

[93] S. Borzsony, D. Kossmann, and K. Stocker, "The skyline operator," in *Proc. ICDE*, Apr. 2001, pp. 421–430.

[94] S. Din, A. Ahmad, A. Paul, M. M. U. Rathore, and J. Gwanggil, "A cluster-based data fusion technique to analyze big data in wireless multi-sensor system," *IEEE Access*, vol. 5, pp. 5069–5083, 2017.

[95] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 10, pp. 1345–1359, Oct. 2010.

[96] S. Jiang, X. Qian, T. Mei, and Y. Fu, "Personalized travel sequence recommendation on multi-source big social media," *IEEE Trans. Big Data*, vol. 2, no. 1, pp. 43–56, Mar. 2016.

[97] S. Jwa, Ü. Ozguner, and Z. Tang, "Information-theoretic data registration for UAV-based sensing," *IEEE Trans. Intell. Transp. Syst.*, vol. 9, no. 1, pp. 5–15, Mar. 2008.

[98] S. K. Barai, "Data mining applications in transportation engineering," *Transport*, vol. 18, no. 5, pp. 216–223, 2003.

[99] S. Kaisler, F. Armour, J. A. Espinosa, and W. Money, "Big data: Issues and challenges moving forward," in *Proc. 46th Hawaii Int. Conf. Syst. Sci. (HICSS)*, Jan. 2013, pp. 995–1004.

[100] S. Khan, M. Shiraz, A. W. A. Wahab, A. Gani, Q. Han, and Z. B. A. Rahman, "A comprehensive review on adaptability of network forensics frameworks for mobile cloud computing," *Sci. World J.*, vol. 2014, Jul. 2014, Art. no. 547062.

[101] S. Madden, "From databases to big data," *IEEE Internet Comput.*, vol. 16, no. 3, pp. 4–6, May/Jun. 2012.

[102] S. Sagiroglu and D. Sinanc, "Big data: A review," in *Proc. Int. Conf. Collaboration Technol. Syst. (CTS)*, May 2013, pp. 42–47.

[103] S. Shekhar, "Spatial big data challenges," in *Proc. Tuts. ARO/NSF Workshop Big Data Large, Appl. Algorithms*, USA, 2012.

[104] S. Sun and C. Zhang, "The selective random subspace predictor for traffic flow forecasting," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 2, pp. 367–373, Jun. 2007.

[105] S. Suthaharan, "Big data classification: Problems and challenges in network intrusion prediction with machine learning," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 41, no. 4, pp. 70–73, 2014.

[106] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.

[107] S. Yu, M. Liu, W. Dou, X. Liu, and S. Zhou, "Networking for big data: A survey," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, pp. 531–549, 1st Quart., 2016.

[108] S.-N. Orzen, M. Stratulat, S. Babii, and C. Cosovan, "Routing tables as big data searchable structures for achieving real-time session fault tolerant rerouting," in *Proc. IEEE Int. Symp. Comput. Intell. Inform. (CINTI)*, Nov. 2016, pp. 000261–000264.

[109] T. Baker, B. Al-Dawsari, H. Tawfik, D. Reid, and Y. Ngoko, "GreeDi: An energy efficient routing algorithm for big data on cloud," *Ad Hoc Netw.*, vol. 35, pp. 83–96, Dec. 2015.

[110] T. H. Davenport, *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities*. HBR Press, 2014.

[111] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer-Verlag, 2008.

[112] T. Kraska, "Finding the needle in the big data systems haystack," *IEEE Internet Comput.*, vol. 17, no. 1, pp. 84–86, Jan./Feb. 2013.

[113] U. Mir and A. Munir, "An adaptive handoff strategy for cognitive radio networks," *Wireless Netw.*, vol. 24, no. 6, pp. 2077–2092, 2017, doi: 10.1007/s11276-017-1455-8.

[114] U. Mir and L. Nuaymi, "LTE pricing strategies," in *Proc. IEEE VTC-Spring*, Jun. 2013, pp. 1–6.

[115] V. Ceikute and C. S. Jensen, "Vehicle routing with user-generated trajectory data," in *Proc. IEEE Int. Conf. Mobile Data Manage. (MDM)*, Jun. 2015, pp. 14–23.
[116] V. Mayer-Schönberger and K. Cukier, *Big Data: A Revolution That Will Transform How we Live, Work, and Think*. Boston, MA, USA: Houghton Mifflin Harcour, 2013.
[117] V. Mehta and I. Chana, "Urban traffic state estimation techniques using probe vehicles: A review," in *Computing and Network Sustainability*. Singapore: Springer, 2017, pp. 273–281.
[118] W. Luo, H. Tan, L. Chen, and L. M. Ni, "Finding time period-based most frequent path in big trajectory data," in *Proc. SIGMOD*, 2013, pp. 713–724.
[119] W. Xia, P. Zhao, Y. Wen, and H. Xie, "A survey on data center networking (DCN): Infrastructure and operations," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, pp. 640–656, 1st Quart., 2017.
[120] Y. Lin, P. Wang, and M. Ma, "Intelligent transportation system (ITS): Concept, challenge and opportunity," in *Proc. IEEE Int. Conf. High Perform. Smart Comput.*, Beijing, China, May 2017, pp. 167–172.
[121] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: A deep learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 865–873, Apr. 2015.
[122] N. Mastronarde, V. Patel, J. Xu, L. Liu, and M. van der Schaar, "To relay or not to relay: Learning device-to-device relaying strategies in cellular networks," *IEEE Trans. Mobile Comput.*, vol. 15, no. 6, pp. 1569–1585, Jun. 2016.
[123] X. Lu, C. Wang, J.-M. Yang, Y. Pang, and L. Zhang, "Photo2Trip: generating travel routes from geo-tagged photos for trip planning," in *Proc. Int. Conf. Multimedia*, 2010, pp. 143–152.
[124] X. Qian, H. Feng, G. Zhao, and T. Mei, "Personalized recommendation combining user interest and social circle," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 7, pp. 1763–1777, Jul. 2014.
[125] X. Wu and X. Zhu, "Mining with noise knowledge: Error-aware data mining," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 38, no. 4, pp. 917–932, Jul. 2008.
[126] Y. Jin, J. Dai, and C.-T. Lu, "Spatial-temporal data mining in traffic incident detection," in *Proc. SIAM Conf. Data Mining, Workshop Spatial Data Mining*, Bethesda, MD, USA, 2006, pp. 1–5.
[127] Y. Lyu, C.-Y. Chow, R. Wang, and V. C. S. Lee, "Using multi-criteria decision making for personalized point-of-interest recommendations," in *Proc. 22nd ACM SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst. (SIGSPATIAL)*, 2014, pp. 461–464.
[128] Y. Shi, P. Serdyukov, A. Hanjalic, and M. Larson, "Personalized landmark recommendation based on geotags from photo sharing sites," in *Proc. 5th AAAI Conf. Weblogs Social Media*, vol. 11, 2011, pp. 622–625.
[129] D. K. Sharma, S. K. Dhurandher, I. Woungang, R. K. Srivastava, A. Mohananey, and J. J. P. C. Rodrigues, "A machine learning-based protocol for efficient routing in opportunistic networks," *IEEE Syst. J.*, vol. 12, no. 3, pp. 2207–2213, Sep. 2018.
[130] Z. Zhou, T. Lin, K. Thulasiraman, and G. Xue, "Novel survivable logical topology routing by logical protecting spanning trees in IP-over-WDM networks," *IEEE/ACM Trans. Netw.*, vol. 25, no. 3, pp. 1673–1685, Jun. 2017.
[131] A. Poorthuis *et al.*, *Using Geotagged Digital Social Data in Geographic Research*. Rochester, NY, USA: Geography Social Science Research Network, Oct. 2014.
[132] G. Bello-Orgaz, J. J. Jung, and D. Camacho, "Social big data: Recent achievements and new challenges," *Inf. Fusion*, vol. 28, pp. 45–59, Mar. 2016.
[133] B. Pang and L. Lee, *Opinion Mining and Sentiment Analysis (Survey)*, 2nd ed. New York, NY, USA: Cornell Univ. Press, 2008.
[134] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014.
[135] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," *Knowl.-Based Syst.*, vol. 89, pp. 14–46, Nov. 2015.
[136] H. Chen, R. H. L. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact," *MIS Quart.*, vol. 36, no. 4, pp. 1165–1188, Dec. 2012.
[137] J. LaRiviere, P. McAfee, J. Rao, V. K. Narayanan, and W. Sun. (May 2016). Where predictive analytics is having the biggest impact. Harvard Business Review. [Online]. Available: https://hbr.org/2016/05/where-predictive-analytics-is-having-the-biggest-impact
[138] S. Erevelles, N. Fukawa, and L. Swayne, "Big data consumer analytics and the transformation of marketing," *J. Bus. Res.*, vol. 69, no. 2, pp. 897–904, 2016.
[139] J. Barney, "Firm resources and sustained competitive advantage," *J. Manage.*, vol. 17, no. 1, pp. 99–120, 1991.
[140] M. van Rijmenam. (Sep. 2017). Southwest airlines uses big data to deliver excellent customer service. Data Floq. [Online]. Available: https://datafloq.com/read/southwest-airlines-uses-big-data-deliver-excellent/371
[141] C. Duhigg. (Feb. 2012). How companies learn your secrets. New York Times Magazine. [Online]. Available: http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html
[142] J. Li, X. Li, and B. Zhu, "User opinion classification in social media: A global consistency maximization approach," *Inf. Manage.*, vol. 53, no. 8, pp. 987–996, 2016.
[143] Z. Zhang, X. Li, and Y. Chen, "Deciphering word-of-mouth in social media: Text-based metrics of consumer reviews," *ACM Trans. Manage. Inf. Syst.*, vol. 3, no. 1, pp. 1–23, Apr. 2012.
[144] J. Qi, Z. Zhang, S. Jeon, and Y. Zhou, "Mining customer requirements from online reviews: A product improvement perspective," *Inf. Manage.*, vol. 53, no. 8, pp. 951–963, 2016.
[145] J. Hea, H. Liu, and H. Xiong, "*SocoTraveler*: Travel-package recommendations leveraging social influence of different relationship types," *Inf. Manage.*, vol. 53, no. 8, pp. 934–950, 2016.
[146] Z. Zhou *et al.*, "A method for real-time trajectory monitoring to improve taxi service using GPS big data," *Inf. Manage.*, vol. 53, no. 8, pp. 964–977, 2016.
[147] T. B. Murdoch and A. S. Detsky, "The inevitable application of big data to health care," *JAMA*, vol. 309, no. 13, pp. 1351–1352, 2013.
[148] S. Kumar, W. Nilsen, M. Pavel, and M. Srivastava, "Mobile health: Revolutionizing healthcare through transdisciplinary research," *Computer*, vol. 46, no. 1, pp. 28–35, Jan. 2013.
[149] M. A. Barrett, O. Humblet, R. A. Hiatt, and N. E. Adler, "Big data and disease prevention: From quantified self to quantified communities," *Big Data*, vol. 1, no. 3, pp. 168–175, 2013.
[150] J. Wu, H. Li, S. Cheng, and Z. Lin, "The promising future of healthcare services: When big data analytics meets wearable technology," *Inf. Manage.*, vol. 53, no. 8, pp. 1020–1033, 2016.
[151] I. Husain. (Sep. 2014). Researchers show Google glass can calculate heart rate and respiratory rate. iMedicalApps. [Online]. Available: https://www.imedicalapps.com/2014/09/google-glass-calculate-heart-rate-respiratory-rate/
[152] N. Sultan, "Reflective thoughts on the potential and challenges of wearable technology for healthcare provision and medical education," *Int. J. Inf. Manage.*, vol. 35, no. 5, pp. 521–526, 2015.
[153] N. Sultan, "Making use of cloud computing for healthcare provision: Opportunities and challenges," *Int. J. Inf. Manage.*, vol. 34, no. 2, pp. 177–184, 2014.
[154] M. F. Goodchild and J. A. Glennon, "Crowdsourcing geographic information for disaster response: A research frontier," *Int. J. Digit. Earth*, vol. 3, no. 3, pp. 231–241, 2010.
[155] G. Cervone, E. Sava, Q. Huang, E. Schnebele, J. Harrison, and N. Waters, "Using Twitter for tasking remote-sensing data collection and damage assessment: 2013 Boulder flood case study," *Int. J. Remote Sens.*, vol. 37, no. 1, pp. 100–124, 2016.
[156] M. Goodchild, "Citizens as sensors: The world of volunteered geography," *GeoJournal*, vol. 69, no. 4, pp. 211–221, 2007.
[157] S. Dashti, L. Palen, M. P. Heris, K. M. Anderson, T. J. Anderson, and S. Anderson, "Supporting disaster reconnaissance with social media data: A design-oriented case study of the 2013 Colorado floods," in *Proc. 11th Int. ISCRAM Conf.* State College, PA, USA: Univ. Park, 2014.
[158] J. Howe, *Crowdsourcing: Why the Power of the Crowd is Driving the Future of Business*, 1st ed. New York, NY, USA: Crown Publishing Group, 2008.
[159] Z. Xu *et al.*, "Crowdsourcing based description of urban emergency events using social media big data," *IEEE Trans. Cloud Comput.*, to be published.
[160] J. P. de Albuquerque, B. Herfort, A. Brenning, and A. Zipf, "A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management," *Int. J. Geograph. Inf. Sci.*, vol. 29, no. 4, pp. 667–689, 2015.
[161] R. Ranjan, J. Phengsuwan, P. James, S. Barr, and A. van Moorsel, "Urban risk analytics in the cloud," *IT Prof.*, vol. 19, no. 2, pp. 4–9, Mar./Apr. 2017.

[162] P. Goel, L. Kulik, and K. Ramamohanarao, "Optimal pick up point selection for effective ride sharing," *IEEE Trans. Big Data*, vol. 3, no. 2, pp. 154–168, Jun. 2017.
[163] B. J. Jansen, M. Zhang, K. Sobe, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 60, no. 11, pp. 2169–2188, Nov. 2009.
[164] A. Bharadwaj, O. A. El Sawy, P. A. Pavlou, and N. V. Venkatraman, "Digital business strategy: Toward a next generation of insights," *Manage. Inf. Syst.*, vol. 37, no. 2, pp. 471–482, 2013.
[165] Y. Qu, S. Yu, L. Gao, and J. Niu, "Big data set privacy preserving through sensitive attribute-based grouping," in *Proc. IEEE ICC*, May 2017, pp. 1–6.
[166] S. Yu, "Big privacy: Challenges and opportunities of privacy study in the age of big data," *IEEE Access*, vol. 4, pp. 2751–2763, 2016.
[167] R. Y. K. Lau, J. L. Zhao, G. Chen, and X. Guo, "Big data commerce," *Inf. Manage.*, vol. 53, no. 8, pp. 929–933, 2016.
[168] H. Yao, M. Xiong, D. Zeng, and J. Gong, "Mining multiple spatial–temporal paths from social media data," *Future Gener. Comput. Syst.*, vol. 87, pp. 782–791, Oct. 2017.
[169] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.
[170] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and POIs," in *Proc. 18th Int. Conf. Knowl. Discovery Data Mining (ACM SIGKDD)*, 2012, pp. 186–194.
[171] D. Sui and M. Goodchild, "The convergence of GIS and social media: challenges for GIScience," *Int. J. Geograph. Inf. Sci.*, vol. 25, no. 11, pp. 1737–1748, 2011.
[172] N. Godbole, M. Srinivasaiah, and S. Skiena, "Large-scale sentiment analysis for news and blogs," in *Proc. Int. Conf. Weblogs SM (ICWSM)*, 2007, pp. 1–4.
[173] L.-W. Ku, Y.-T. Liang, and H.-H. Chen, "Opinion extraction, summarization and tracking in news and blog corpora," in *Proc. Spring Symp. Comput. Approaches Anal. Weblogs (AAAI-CAAW)*, 2006, pp. 1–8.
[174] A. Alsayat and H. El-Sayed, "Social media analysis using optimized K-Means clustering," in *Proc. IEEE Int. Conf. Softw. Eng. Res., Manage. Appl.*, Jun. 2016, pp. 61–66.
[175] R. Kaur and S. Singh, "A survey of data mining and social network analysis based anomaly detection techniques," *Egyptian Inform. J.*, vol. 17, no. 2, pp. 199–216, 2015.
[176] W. Shi, A. Zhang, X. Zhou, and M. Zhang, "Challenges and prospects of uncertainties in spatial big data analytics," *Ann. Amer. Assoc. Geographers*, to be published, doi: 10.1080/24694452.2017.1421898.
[177] C. A. Ferrero, L. O. Alvares, and V. Bogorny, "Multiple aspect trajectory data analysis: research challenges and opportunities," in *Proc. GEOINFO*, 2016, pp. 56–67.

USAMA MIR received the B.S. degree (Hons.) in computer engineering from the Balochistan University of IT, Engineering and Management Sciences, Pakistan, in 2006, and the master's and Ph.D. degrees in computer science from the Troyes University of Technology, France, in 2008 and 2011, respectively. He was a Post-Doctoral Fellow with Telecom Bretagne, France, from 2011 to 2012. He was the Head of the Electronics Engineering Department, Iqra University Islamabad, Pakistan, from 2012 to 2015. He is currently an Assistant Professor with Saudi Electronic University, Saudi Arabia. His research interests include big data analysis, resource allocation and handoff management in cognitive radio systems, wireless communications and networking, Markov chains, next-generation networks, and multi-agent systems.

UBAID ABBASI received the M.S. degree from SUPELEC, Rennes, France, in 2008, and the Ph.D. degree from the University of Bordeaux, France, in 2012. He was a Senior Research Fellow with the University of Quebec, Canada, where he was involved in a project funded by Ericsson, Canada. He is currently an Assistant Professor with the Department of Computer Science, GPRC, AB, Canada. His research interests include inter-container communications, data center communication issues, device-to-device communication in the next-generation 5G networks, wireless communications, and big data analysis.

YANG YANG (M'13) received the master's and Ph.D. degrees from the School of Telecommunications Engineering, Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2012 and 2015, respectively. From 2015 to 2017, he was a Post-Doctoral Researcher with the Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing. He is currently an Assistant Professor with the Center for Data Science, BUPT. His research interests include AI-based wireless networks, machine learning, and 5G ultra-dense networks.

ZEESHAN AHMED BHATTI received the B.S. degree in computer science and the M.S. degree in engineering management in Pakistan, and the M.S. and Ph.D. degrees in management information systems from IAE Aix en Provence, Université Aix Marseille, France. He was with universities in Pakistan and multinational firms like Ericsson, Nortel, and Siemens. He is currently an Assistant Professor with the Faculty of Economics and Administration, King Abdulaziz University, Saudi Arabia. His has authored in international peer-reviewed journals and presented in various international conferences. His research interests are social media in business, knowledge management, and electronic commerce.

TALHA MIR received the B.S. degree in electronic engineering from the Balochistan University of IT, Engineering and Management Sciences (BUITEMS), Pakistan, in 2007, and the master's degree from the University of Bradford, England, in 2011. He is currently pursuing the Ph.D. degree with Tsinghua University, Beijing, China. He is also an Assistant Professor with BUITEMS. His research interests include resource wireless communications and networking, next-generation networks, massive multi-in multi-out, mm-waves, and spatial movements.