D$^2$SC: Data-Driven Smarter Cities

Chao Chen†, Daqing Zhang†, and Bin Guo‡
† CNRS UMR 5157 SAMOVAR, Institut Mines-TELECOM/TELECOM SudParis, Evry 91011, France
‡ School of Computer Science, Northwestern Polytechnical University, Xi’an 710072, China

Abstract—Data revealing lots of facets about cities are becoming increasingly available, which provides an unprecedented opportunity to get and understand a whole picture of how cities work. In this paper, we first give a brief introduction to the study of data-driven smarter cities (D$^2$SC); second, we introduce and classify some common data available about cities; third, we present four main research objects and features in D$^2$SC, and survey the common research issues dealing with each objects; finally, we propose a general framework and discuss the research challenges.

Index Terms—data-driven, smarter cities; features; framework; challenges

I. INTRODUCTION

Cities, just like human bodies, act as “systems of systems”, involving the interaction of a variety of private and public systems or networks [10]. These systems operate in parallel, and they are interconnected and subject to non-linear dynamics. Nowadays, cities are becoming more and more crowded since enormous growing people are continuing to move into the cities. By 2050, an estimated 70% of the world’s population will live in cities up from 13% in 1900. For China only, about 300 million rural inhabitants will move into cities over the next 15 years, hence the world is featured by urbanization. The world is facing numerous social and environmental city problems which are emerging along with the process of urbanization, such as traffic congestions, the rise of CO$_2$ emissions, the imbalance between energy supply and demand. Therefore, to build cities in which the social, commercial, and environmental needs are balanced whilst the available resources are optimized is a global imperative [9]. Such cities are also known as smarter cities. Cities in world have experienced a process of “cities→digital cities→smarter cities”, which is illustrated in Fig. 1. Also we can see that cities are becoming larger in space with the time. The focus of digital cities is to connect communities by combining broadband communications infrastructures, which lays a solid foundation for building smarter cities.

By the “systems of systems” view, the very fundamental to create a smarter city is to understand how a city functions in a holistic and comprehensive view. With the increasingly wide deployment of sensors, the advent of digital communication and computational technologies, many facets about cities, including people’s movements (e.g. where and when people leave for) and interactions with the city resources (e.g. how people interact with city infrastructures) can be easily digitized and collected [20]. We have entered an era where massive sources of urban data are becoming increasingly available. Such rich data available to cities come from a variety of city systems as well as their interactions, providing new world’s natural resources to understand cities as a whole and help to make more informed decisions about cities. Data-driven smarter cities (D$^2$SC) is the study of turning the raw data into intelligence hidden behind the data itself by applying data-driven approaches/tools to: 1) design, planning and evaluation of city infrastructures to improve the living quality of people; 2) minimize the carbon footprint and meet pollution reduction targets to build an environment-friendly society. Taking data-driven approaches not only could save the money cost of municipal governments and local organizations (i.e. economy), but also could identify potential city problems timely (i.e. near real-time response).

The great potential of D$^2$SC can already be seen in recent studies [2], [3], [4], [7], [8], [15], [21]. For instance, there has been a large body of work on developing data-driven approaches/tools to deal with traffic flows related problems, such as driving directions navigation, estimating the travel time and Origin-Destination (OD) flows, and detecting flawed urban planning [1], [5], [18], [22]. With the taxi GPS data, Yuan et al. [18] present T-Drive system to navigate driving directions by leveraging the intelligence learned from the taxi drivers. Chen et al. [5] proposed B-Planner to provide cost-effective and environment friendly transport to citizens at night. Still other works have sought data-driven approaches to characterize the functionality of a region using social media [14], [17].

In this paper, we aim to: 1) introduce the research objects and the features of D$^2$SC; 2) propose a general framework for

---

2http://cities.media.mit.edu
TABLE I

<table>
<thead>
<tr>
<th>Data sets</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobile phone data</td>
<td>the information about the connected cell tower ID and the calling/texting</td>
</tr>
<tr>
<td>Vehicular GPS</td>
<td>the location and time information about the moving vehicular</td>
</tr>
<tr>
<td>Shared-bicycle usage data</td>
<td>the information about the bicycle station ID and the time when people rent</td>
</tr>
<tr>
<td>Photo-sharing data</td>
<td>and return bicycles</td>
</tr>
<tr>
<td>Smarter transportation cards</td>
<td>the photo and time information when people eat/travel</td>
</tr>
<tr>
<td>Check-in data</td>
<td>the bus station ID and time when people gets on and off</td>
</tr>
<tr>
<td>Home water/electric usage data</td>
<td>the information about venue ID and time when people visit</td>
</tr>
<tr>
<td></td>
<td>the information about people usage of water/energy</td>
</tr>
</tbody>
</table>

D²SC and identify the research challenges.

II. DATA AVAILABLE TO CITIES

There are many types of data that cities are producing, captured and collected every second. Different data can reflect different aspects of facets about cities, and many of them are correlated and complementary to each other [12]. According to the interested objects, we can classify data into the following three groups.

- Data about “People” only: This group includes the data about the moving trajectories of individuals, the social networks among people in real life and on-line, the healthcare, and so on. A representative on-line social network is Facebook⁴, in which rich social relationships are kept, and moreover, user’s profile, such as the gender, age, interests, education level, is also maintained for friend recommendation.
- Data about “Resources” only: This group mainly includes the data about the city infrastructures (e.g. bus stations, energy) and environments (e.g. air/water pollution, wastes). Common data sets are the road network data, POI (Points of Interest) data, the bus station network data, the air quality data.
- Data about “People-Resources”: This group of data records the information about how people use/impact the resources. Table I summarize some representative data sets belonging to this group.

III. RESEARCH OBJECTS AND FEATURES

In this section, we will first present the four major research objects of data-driven smarter cities (D²SC), and then followed by the introduction of its research features.

A. Research Objects

The research of D²SC has four main objects, as shown in Fig. 2, that is, people, resource, place and technology respectively, detailed as follows.

1) People: People are moving, interacting among themselves in the space. Besides, people do not live alone and they are connected and belonging to different groups/communities by social relationships. Thus people are featured by mobility and sociality. They are the creators of cities and also the beneficiaries of D²SC.

The common research issues about “People” mainly include the following aspects:

- Human mobility modeling: the human movement follows some physical laws, such as the power law. Researchers often model the mobility by a multi-level view (i.e. individual-group-community-society), subject to different applications. By modeling human mobility, we can synthesize large-scale human mobility traces for the system scalability test, the algorithm performance evaluation, amongst others [11].
- Human mobility patterns discovery/understanding: Studies in this aspect discover human mobility patterns and understand the motivation/intention behind the patterns, such as the phone calling patterns for an individual, based on their historical mobility traces. Again, human mobility patterns can also be in multi-level. By discovering and understanding the human mobility patterns, city resources can be better allocated and distributed.
- Human mobility/activity prediction: Studies in this aspect answer questions, such as what will her do in the next few minutes/hours/days and where will she go in the next few minutes/hours/days.
- Community detection: Studies in this aspect is to detect user communities inside which people are better interacted among themselves. The community can be either from the real social networks (e.g. family), or from the on-line social networks (e.g. Facebook friends). In a much
boarder sense, people who share similar preferences (e.g. visiting similar places) can be also identified as the same community even if they do not know each other.

2) Resource: Resource consists of several elements, such as the infrastructures (e.g. buildings, surveillance video cameras, city roads), energy (e.g. water, fuel), environments (e.g. air, wastes). One key feature of the resource is that it is limited. Besides, the components of the resource are evolving/changing with the time. For instance, the road network is evolving as new road segments are created to add into it, meanwhile, some old road segments will disappear. The air quality in a specific region is changing with the time.

The common research issues about the “Resource” include the following aspects:

- **Resource planning/evaluation**: Studies in this aspect focus on developing data-driven approaches to plan the resource to meet the users’ demands, and evaluate the effectiveness of current resource distribution. For instance, resource planning can guide city planner to solve the questions, such as how many infrastructures (i.e. bus stations, loop sensors) and how to distribute them in a city? By analyzing the city-wide OD flow patterns, new bus routes can be planned to improve the public transportation [5]. We can also examine supply-demand conditions to see whether the current shared-use bicycle stations have been well planned.

- **Traffic/environment monitoring and inference**: Studies in this aspect focus on developing advanced approaches to monitor and inference traffic/air quality of the whole city based on the incomplete observations from limited city road segments and air quality measurement stations.

3) Place: Place is the location information of people and the resource. People and the resource would become more meaningful if the additional location information is associated, and many location-based services can be also enabled. Nowadays, many sensors and devices, such as WiFi, GPS, mobile phone, have been used to detect and report the location information. The reported location information can be a physical location (i.e. the longitude and latitude coordinates) or a symbolic location (e.g. the Louvre Museum).

The common research issues about the “Place” include the following aspects:

- **Advanced localization methods**: Studies in this aspect develop more accurate, energy-efficient, and equipment-less methods to identify the user’s location, according to different applications.

- **Location annotation**: Studies on this aspect convert raw meaningless location into meaningful logical location with semantic annotation to understand the user’s intention better [19].

- **Location-based business/service**: Studies in this aspect mainly include location-based recommendation services, itinerary guidance, driving directions navigation.

4) Technology: Technology digitalizes, collects, transits, stores and analyzes the observations about people (i.e. the number of visiting places and orders, the visiting frequency), the resource (i.e. the air quality level of a region at a certain time), the place and their interactions as well (i.e. people check-in a restaurant through Foursquare). Technology bridges the gap among them, and moreover, the technology has also revolutionized the way in which people interact with the resource. It is no exaggeration that cities now embrace technology as a strategic instrument for shaping their future.

The common research issues about the “Technology” include:

- **Advanced sensing technologies**: Studies in this aspect mainly include sensing networking, mobile crowd sensing [6].

- **Advanced data transmission technologies**: Research in this aspect studies how to devise effective and efficient routing protocols and algorithms to transit the data into the data center.

- **Advanced data management technologies**: Studies in this aspect include data storage, integration, and query.

- **Advanced data mining and visualization techniques**: Data mining techniques can be broadly classified into five groups: clustering, classification, regression, ranking, and optimization. Different data mining techniques are designed addressing different research problems. For example, with a clustering algorithm of merging-splitting process, bus stops where lots of people visit can be identified [5]. Besides, data visualization provides a more intuitive way to see underlying patterns in data.

B. Research Features

The research of the D²SC has the following features:

- **Using data to optimize cities**: This is the most salient feature of D²SC, however, what applications can be enabled are highly depend on the inherent characteristics of the available data sets.

- **“Big Data”**: The big data is known as its four dimensions: volume, velocity, variety, and veracity. As discussed in Section II, data about cities presents the “big data” features. Different kinds of data are growing in every second from multiple sources, thus the scale of data is huge (i.e. volume). They are also highly heterogeneous and diverse in form (i.e. variety), such as the texts, images. Take the taxi GPS data as an example, taxi reports its locations frequently and send them into the server in streams (i.e. variety). Moreover, since the sampling rate of the taxi GPS data is about one minute, the trajectory between two consecutive reported locations is uncertain (i.e. veracity).

- **“Citizen” as an additional resource**: As a result of the ubiquitousness and powerfulness of mobile devices, citizens also act as data contributors to observe and diagnose the city situations, enabling to get a more complete and deep picture. For instance, with their mobile phones, users can measure and upload the surrounding air quality conditions. With the help of data contributed by individuals in different city regions, a full city-wide picture
of air quality can be puzzled. Furthermore, the reasons why the air quality is abnormal (i.e. goes worse/better suddenly) in a region can be also diagnosed through mining user generated contents (e.g. twitter). Another interesting research issue in this line is how to design incentive mechanisms appropriately to attract more users to participate [16].

- **Interdependence**: From the “systems of systems” point of view, each system inside the city is interdependent and interconnected. First of all, output of a system also serves as the input for the other systems, and plays impacts on them. For instance, city planners should place a suitable number of bus stations where lots of people frequently visited when designing a new bus line, meanwhile, more people than before would be gathered along the new bus line after it is open. Secondly, improvements for a system only may fail to lead to a smarter city eventually. For example, to alleviate the traffic jams, many cities in China are introducing a new transportation mode (i.e. Metro). This approach has been proved to improve the traffic significantly, however, can be costly and demanding for the user of already very limited land resources. The feature requires to treat cities and make better decisions to build smarter cities as a whole.

IV. RESEARCH FRAMEWORK AND CHALLENGES

In this section, based on the elaboration of the research objects and features of D^2SC, we propose a reference framework to illustrate the key functional blocks, and followed by the introduction of key challenges.

A. Framework

Figure 3 shows the proposed framework, which is a closed-loop system and composed by four layers: data acquisition, data integration and management, data analytics, and applications.

- **Data acquisition**: The first layer collects all kinds of data growing in every second from different sources through networking to data centers. Nodes responsible for transmitting data can either be static or mobile. For instance, during a mobile crowd sensing task, the nodes are moving and selected in an ad-hoc or opportunistic way.

- **Data integration and management**: The second layer is in charge of integrating and managing such big data sets in an efficient way, which is extremely challenging. More difficulty, some data come to the center in a stream manner which requires the integration and management techniques can handle them in real time. Last bus not least, privacy issues should be also considered before data publishing.

- **Data analytics**: This layer applies reasoning, data mining, data visualization to turn the collected low-level, single-modality sensing data to the expected “intelligence”. A preliminary is to capture and generalize user requirements, then mine “intelligence” to support a variety of urban services. To support complex applications and services (e.g. traffic diagnose), “intelligence” should be fused from diverse but complementary data sets in an integrated level (e.g. the bus GPS data, event data, user generated content).

- **Applications**: This layer includes many potential applications and services, varying from living, working, transportation, healthcare, education, which can be roughly categorized into 7 groups, as shown in Fig. 4. For example, “Smarter Transportation” mainly concerns with the provision of traveller information (e.g. the information about the real-time traffic and when will the next bus/train come), the recommendation of itinerary and driving routes, exploring the “intelligence” mined from relevant data sets. An important shared function among all applications is the user interface. A friend user interface is also of great importance for data visualization and effective interaction between machines and people to facilitate the user in decision-making process.

B. Key Research Challenges

The research of D^2ST shares all the challenges as the research of “Big Data”. Besides, it also suffers from the following challenges:

- **Data quality and trust**: One feature of the research is that it explores the data contributed by citizens, however, this kind of data would have distinct quality and credibility as contributors may have different devices, backgrounds and knowledge. Thus how to clean and remove erroneous sensing observations, and identify false data automatically before data analytics is very challenging.
• **Data integration:** Although a large number of data sets from various different domains became available, they come in diverse formats, have little semantic meaning, and do not link to each other. The challenges are how to convert the unstructured data into well-structured, semantic data and link them. Furthermore, identification of heterogeneous data sets that are closed related, as a part of data integration, is also very challenging [13]. In addition, fused intelligence from multiple data sources in an integrated level would aid better decision-making in a number of ways.

• **Data sparseness:** Due to the limitation of resources, data are often sparsity both in *time and space*. For instance, to save energy, sensors often send sensing results to centers at irregular possibly large time intervals. Likewise, only a limited number of loop sensors installed in main road segments to measure their traffic conditions. However, building a smarter city requires a full city-wide data, thus data inference based on the sparsity data is a pre-condition, which is difficult. To overcome the data sparseness problem, many algorithms have been devised. The popular methods are based on the collaborative filter, whose basic idea is to explore the similarity among users or places. There are also some methods leveraging the complementary information from the other relevant data sets.

• **Data biases:** Available data sets can be biased and over represented as they are based on a certain socio-economic group. It is very challenging to correct them.

• **Privacy issues:** Data can contain sensitive information, such as the people’s preferences, social relationships, address, bank cards number, income. The disclosure of such information may result in disastrous consequences. Privacy issues should be concerned during the data collection, integration, pushing, and sharing. The challenges are to keep data fidelity for applications while protecting privacy.

V. CONCLUSION

With the mass-proliferation of smart-phones and the wide-deployment of various of sensors, rich sources of large scale data revealing how people and their impactions on city infrastructures and environments are becoming increasingly easy to obtain, enabling us to get a full understanding of cities. In this paper, we have presented the study of smarter cities by applying data-driven approaches, which has been proven to be an economic way and also have timely response. A reference framework about from a variety of city “data” to “intelligence” is also proposed. Finally, we discussed the research features and challenges we should handle.

REFERENCES


