A Crowd-Sourced Data Based Analytical Framework for Urban Planning

Li Dong, Long Ying

Abstract Aimed at the challenges faced by the current urban development and urban planning, along with the research opportunities brought by "big data," this paper proposes an analytical framework based on crowd-sourced data for urban planning by reviewing related literature and practice. The framework is mainly oriented towards three major requirements of analysis in urban planning: the physical spaces, the user communities, and the social relationships. This analytical framework can be regarded as a preliminary attempt for future data-intensive applications in urban planning and assessment.

Keywords location-based services (LBS); crowd-sourced; natural language processing; quantitative urban study

1. Introduction

1.1 Challenges for urban development and urban planning

After 30 years of rapid urban development, China currently has an urbanization rate of more than 50%. Many negative impacts of urbanization, i.e., so-called "urban diseases," are emerging, including traffic congestion, excessive population concentration, heavy consumption of resources, environmental pollution, poor safety and disaster prevention, and so on. Urban diseases are testing the capabilities of urban management and sustainability, with a tendency of spreading out from mega cities to less developed small and medium-sized cities. Since the *1996 Istanbul Declaration on Human Settlements* proposed a statement of "making human settlements safer, healthier, and more livable, equitable, sustainable and productive," the gap between the goal and the reality has become even larger, and the living quality of urban residents has faced serious challenges.

Meanwhile, as a leading player, urban planning itself also faces many threats. Since the conditions and status of cities have changed significantly and are becoming extremely complex, the effectiveness of traditional tools for urban planning such as technical standards and analytical methods has been declining. Planners' abilities of analyzing, diagnosing, and assessing the status of urban development are doubted, let alone their abilities to guide future development, to implement proactive solutions to issues in real world, or to enhance the feasibilities of targets. One of the reasons is the insufficiency of traditional data and analysis. Planners have to extract information based on a relatively small amount of data and try to seek an overall conclusion. In other words, the mode of analysis requires an effective transition from fragmented, low-frequency statistics to a complex overview picture of a city, as well as a shift from a rough aggregation of figures to a finer profiling of individuals (Zhang, 2014).

1.2 The "big data" wave

With the booming of information and communications technology (ICT), incredible amounts of data are produced and available in our cities and on our planet, via various chips, sensor networks, positioning systems, mobile communications, and high-performance computing and storing technologies. Urban daily life, such as transportation and recreation, has also been impacted by the evolving ICT. For example, Baidu.com processes 6 billion search requests every day; over 500 million people talk via WeChat app and compose over 100 billion relationships online; the bus-pass card in Beijing is used up to 20 million times per day, and so on. Human and various types of operation sensors will produce more and more data. According to the estimates from the white paper of the International Data Corporation (IDC) (Gantz and Reinsel, 2009), for every 18 months the volume of new data is equal to the sum of all data in the past, while the total amount of data generated each year will reach 40 ZB by 2020.

To respond to the "big data" wave and further reveal its impacts on cities and human society, the academic community has carried out a considerable amount of research works, represented by two special issues published respectively by *Nature* (2008) and *Science* (2011). Data deposits will be increasing gradually from the level of GB (GigaByte) to PB (PetaByte) and EB (ExaByte), while effective meta-analyses on these data with complex structures and huge amounts, especially obtained from the urban residents' life behaviors as well as social and public opinions online, are valuable for inferring contexts and trends considering the overall process of urban evolution. The goal of research is to realize an optimal development by exploring the dynamics of natural and social laws underlying the urban development in the past and present.

1.3 Meaningful crowd-sourced data

Besides the sensors and other monitoring data from the Internet of things (IOT), the most valuable contributions are still from the public. People's ubiquitous activities generate a lot of data each day, such as track records, communications via mobile phones, consumption, web browsing, and online social networking. All of the data reflects the interaction between their producers and the surrounding environment or even the whole city. Thus the so-called crowd-sourced data, defined by the way it is generated, equally become a valuable source of information for observing and evaluating a city.

The concept of "social sensing" (Yu et al., forthcoming) has pointed out that human individuals have constituted a new sensor network in the city in addition to the traditional sensing sources such as satellites. Thanks to the popularization of smart devices such as smart phones and tablets, crowd-sourced data contains a large amount of time and location information in these logs. These kinds of crowd-sourced geographic data are important for urban study and planning. Representative types of data include not only the GPS route data (such as OpenStreeMap, OSM), user collaboratively edited maps (such as Wikimapia), various kinds of social networking messages and check-in data (such as Weibo, 58.com, dianping.com, Twitter, and Facebook), but also a huge amount of data on human activities recorded and observed by the devices (such as cell phone positioning, bus cards, and taxi positions).

Based on these kinds of data, spatio-temporal models on human behaviors in combination with other relevant social factors can provide a useful interpretation for urban planning, public safety, and economic development (Zhou et al., 2013). In particular, in the area of transportation researches, dynamic trajectories formed by continuous points have been greatly explored, supporting a number of laws helping urban management and regulation.

Take urban transportation for example. From the perspective of urban planning, understanding the residents' job-housing distribution is a good start for carrying out effective research. However, the traditional means of questionnaire survey usually waste a lot of time as well as human and financial resources. Most of the survey samples range from hundreds to tens of thousands (Zhou et al., 2013), which is a small part compared to the millions of people living in a city. Researchers have to increase the complexity of modeling, so as to reduce the potential dependence on subjective judgment on parameters. A massive increase in the research samples can improve the representativeness and reliability of the results (Long, Zhang, and Cui, 2012).

Compared with traditional methods, the new quantitative tools for urban planning have a breakthrough in the two aspects: one is to provide a finer dynamic tracing method towards the city in dimensions of both space and time; the second is to improve the logic chain for the interaction between the physical space and the individual/group behaviors for a better interpretation. To achieve this goal, it is necessary to establish an integrated analytical framework based on crowd-sourced data, which is general and can be applied to different types of data. The outputs can be compared, after the clarification of contents, methodology, and indicators of analysis. Important parameters should be filtered and refined to evaluate its capability of answering various problems in urban planning.

2. Analytical framework

Both the policy and technical goals of urban planning should be implemented by regulating the physical built-up environment. However, its ultimate objective is to better serve people's activities in cities. Thus, the analytical framework for urban planning should be applied to three aspects: physical space, user communities, and social relationships.

2.1 Physical space: boundaries, functions, and semantics

Cognition issue of physical space in urban planning have the following questions: Where should the boundary of a city center be delineated? What is the primary function of a region? And how do people think of a site or an area?

For the urbanized areas, Gao (2014) proposed a Densi-Graph model to identify regional boundaries of urban agglomerations through seeking the turn-over point at the density curve from a large number of Points of Interest (POIs). Zhou et al. (2014) utilized mobile phone positioning data of users in Shenzhen to identify the location of city center by drawing three different units of mobile base stations, traffic analysis units (TAZs), and a uniform grid, and discussed the differences of delineation results.

In Shanghai, Niu et al. (2014) used mobile phone users' data to discuss the spatial structure. Combining the location of base stations and users, a kernel density estimation (KDE) method was used to generate the distribution at typical times. The ranking and the major types of urban centers/sub-centers were drawn by clustering their densities and degree of mixing.

For identifying or evaluating the functional type of sites in cities, Guenther et al. (2014) compared the activity pattern of mobile phone users during weekdays and weekends. By spatially clustering the temporal characteristics, they described and verified the differences of various urban sites in Udine of Italy, which supported their planning.

In addition, even for the same physical space, the ways that different users think of and utilize it are also quite different. This is a question urban planners have to answer. The traditional urban planning tends to focus on stipulating the manner of land use and its function, but fails to measure to what degree the stipulation can be implemented. From the technological aspect, the applications of detecting public sentiment via Weibo or hot news are popular, but few of them consider the context of geographic location. Combing the computer-based Natural Language Processing (NLP) tool with location data, ranking the keywords, or extracting the topics occurring on-site, is helpful for understanding any sites in cities from the viewpoint of (physical) "space" to (conceptual) "place" (see Figs. 1 & 2). Some recent researches include semantic analysis of geographic entities by Du et al. (2014) and Duan et al. (2014), while Chen et al. (2014) proposed a spatio-temporal extraction model for topics that were put forward for micro-blogs and names of places.

For the analysis applied to physical space, crowd-sourced data are utilized in the following ways: for density analysis such as KDE and semantic analysis by NPL, to define the boundaries, reinvent cognitive function, and further identify the characteristics by comparing utilization patterns within different time periods. These approaches are significant for understanding the city and evaluating the degree of planning implementation.

2.2 User communities: distribution and mobility

One of the main features in a city is the gathering of population. Thus the spatio-temporal points of users (distribution) (see Fig. 3), and their traces generated by continuous distribution (mobility) are highlighted in the researches of urban and transportation planning. To a certain extent, the spatial patterns of the trajectories of moving objects reflect the intentions and preferences of individuals or groups by their behaviors. Applications using trajectories include network traffic analysis, urban traffic pattern

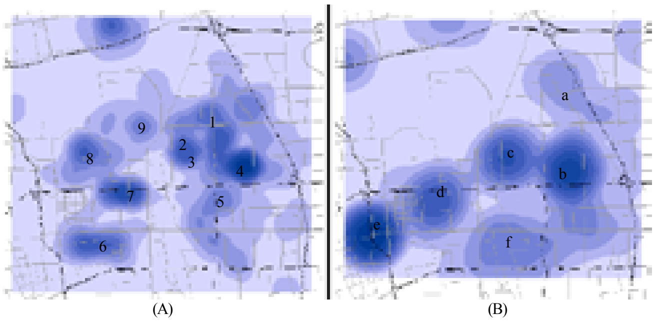

Fig. 1 Difference of semantic recognition to Zhongguancun, Beijing

Note: (A): "university", 1. China University of Mining and Technology; 2. University of Science & Technology Beijing; 3.Beihang University; 4.Beijing Language and Culture University; 5. Tsinghua University; 6. Peking University; 7. University of Chinese Academy of Sciences; 8. Renmin University; (B): "shopping", a. Jinma Tower; b. clothing market; c. Wudaokou; d. Zhongguancun market; e. Wanliu mall; f. Dazhongsi market.

Source: Weibo.

Fig. 2 Topic "congestion" and road network in the city area of Shanghai

Source: Weibo.

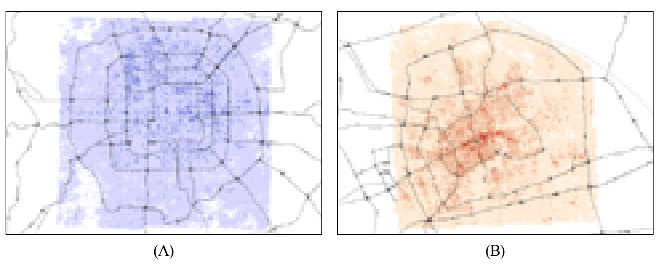

Fig. 3 Spatial distribution of online social platform users in Beijing (A) and Shanghai (B)

Source: Weibo.

analysis, monitoring group events, describing characteristics of human behaviors, and so on.

Using a large set of OSM data generated by users, Jiang and Jia (2011) simulated people's travel behaviors, finding that travel models were mainly affected by the structure of road networks, which could be used for optimizing routes. Zheng and Xie (2010) proposed a recommendation system for popular tourist spots and customized potential places and friends. In Guangzhou, Zhou and Yan (2005) revealed that if the amount of samples reached a certain size, a much more reliable result could be obtained for conducting job-housing balance research. This kind of trajectory-based analysis is meaningful for discovering the spatio-temporal pattern of commuting flows and supporting regulation policies for traffic (Sun, Pan, and Ning, 2008).

Generally, the origin–destination (OD) analysis and statistics, including clustering, mining patterns, and detecting abnormalities, are the main methods of utilizing crowd-sourced data under the topic of distribution and mobility (Lu and Zhang, 2014).

2.3 Social relationships: mapping networks

Compared to the two previous issues, analyzing social relationships is more difficult. One of the reasons is the complexity of the huge amount of data representing the nodes, edges, and behaviors in a network of relationships; while most of the data is non-structured. It is hard to intuitively grasp the whole picture of the network, which remains a challenge for applying large-scale social network analysis in the study of spatial relationships and system structures in cities.

Despite the spatio-temporal feature of social networks in cities, there have been a considerable number of studies investigating the social network in reality or on-line, focusing on indicators such as betweenness centrality or node degree (Wu and Di, 2004; Gao and Wu, 2010). However, the geographic social network may display different laws according to its evolution (Zhou et al., 2013). Experience from traditional network dynamics needs to be carefully examined before its application in regional and urban studies.

From a couple of existing attempts, using mobile phone data for a whole province, Chi et al. (forthcoming) found that nodes with high betweenness centrality in the network were linearly distributed and closely aligned with major transportation corridors, and detected a hierarchy of communities that did or did not correspond to the administrative boundaries. Li et al. (2013) used more than 5.8 million personal files of Weibo users and their social relationship information to construct a virtual social network in major Chinese cities. To respond to the questions of urban systems such as hierarchy or hinterland range, this network of relationships can play an important role besides the traditional network of economy or transportation (see Fig. 4).

3. Application in urban planning practice

3.1 Assessing the degree of mixed land use in Beijing

For the features of land use, Long and Liu (forthcoming) proposed a classification model for the urban parcels in terms of their major functions based on POIs data, development density, and degree of mixed use, which is essential for planning and its post assessment.

In Beijing, regarding the type of land use, the combination of the traditional commanding-style planning system and the recently established market economy is likely to create a mismatch between the plans and the actual implementation. Before the reform and opening-up in 1978, the type of land use was mainly large parcels and with a single purpose due to the state ownership background. In contrast, the recent economic development in Chinese cities has brought about more functional diversity and might be expected to lead to a more mixed development than before. Thus the index of mixed land use proposed by Frank et al. (2004) was introduced to capture the mismatch and to answer the question:

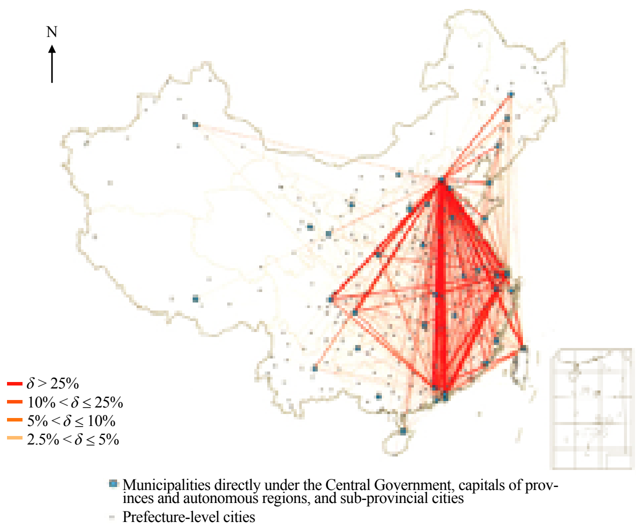

Fig. 4 Inter-city social relationships of China
Source: Weibo. Li Q et al. (2013).

"will the zoning plan and the actual land use reveal different levels of mixed development in Beijing?"

The mixed index (M) of a land lot was calculated as:

$$M = -\sum p_i \ln p_i \quad (i = 1, \cdots, n)$$

where n denotes the number of land use types, and p_i is the proportion of the lot that is characterized by land use type i.

Based on 21,922 planned and 22,027 actual parcels (interpreted from remote sensing images), the authors also used 84,541 POIs and 6,555,529 Weibo check-in records at these POIs to indicate the categories and intensities of each parcel. All of these data were classified into eight land use types according to the standards of urban planning (Long, Gu, and Han, 2012).

The city of Beijing was divided into 2,272 square cells and for each cell the mixed land use index was calculated. Results showed that the overall pattern of land use mixing revealed by each of the different data sources are quite consistent: the degree of mixing remained higher in the center and much lower in the periphery; geographic extents of planned and actual urban activities overlapped largely. However, land use patterns captured by check-ins and POIs data were found more mixed than those measured by parcel-level observations (see Fig. 5). More details of the application are available in Long and Liu (2013).

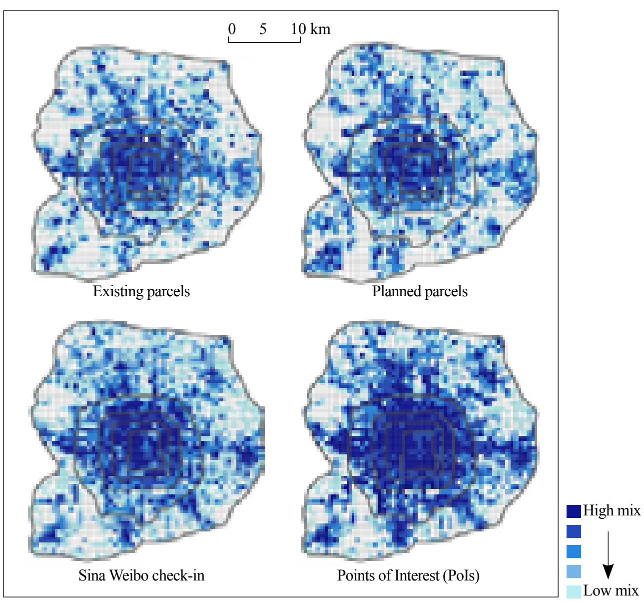

Fig. 5 Visual comparison of mixed land use in Beijing

3.2 Regional integrated analysis during the National Day vacation week

The author conducted a comprehensive analysis on the distribution and flow of people at the national and regional levels as well as semantic mining at selected scenic areas. The raw data contained more than 6.67 million geo-located Weibo records that were extracted from the Application Programming Interface (API) of Weibo during the National Day vacation week in 2014. For the total number, the daily average geo-located Weibo in China was increased by 18% compared with days before the vacation, while the maximun day was 1.02 million and the minimun day was 0.86 million. There was an obvious periodical rise during the vacation week due to users' sharing their holiday experiences (see Fig. 6).

From the aspect of total number, the distribution of geo-located Weibo was similar to the census. Divided by the Aihui-Tengchong Line (ATL), the total number in the eastern part of China was

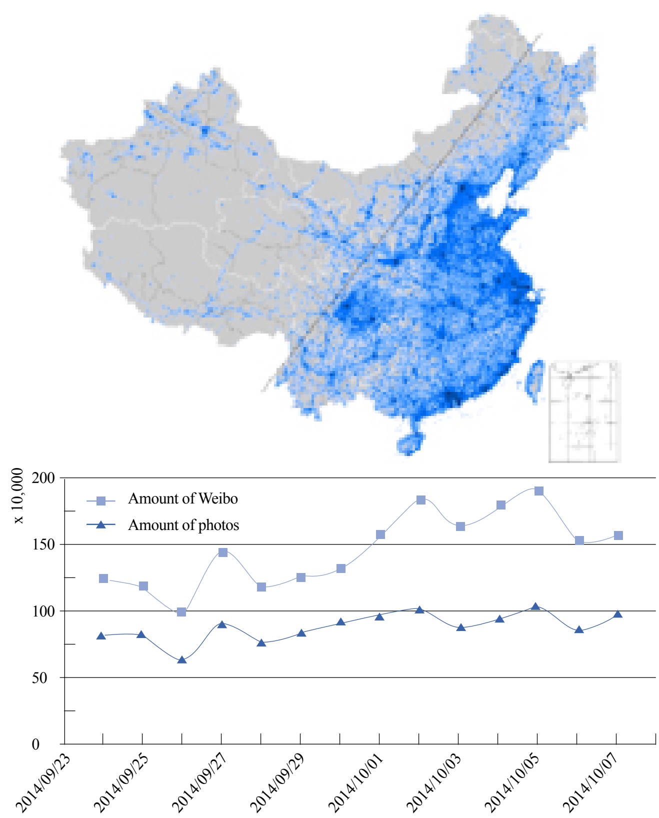

Fig. 6 Spatial distribution of mobile Weibo users (A) and their trend during the National Day vacation week (B)

much higher than that in the western part. Nevertheless, after being compared to the local permanent resident population, the indicator of Weibo per capita revealed the “hot spot” pattern at the county level (see Table 1). Among the top 10 hot counties, five were in Sichuan Province, which was consistent with the official aggregated figures at the provincial level (China National Tourism Administration (CNTA), 2014) (see Fig. 7).

In terms of mobility, the huge amount of points with temporal and spatial information constructed a nationwide network of trajectories. At first the areas of origin and destination for each travel and each user on a city-scale level were defined from the location information of a message or the profile of users. By summing up each single line of arrival and departure we got the national mobility map of social media users during the vacation week (see Fig. 8). A diamond-shaped network formed by tourists, with four apexes of Beijing, Shanghai, Guangzhou, Chengdu and one centroid of Wuhan, could be found visually, covering most urbanized regions of China.

Indicators such as the amount of departure and arrival and their ratio were investigated from different aspects. The top 5 cities receiving most incoming visitors were Beijing, Guangzhou, Shenzhen, Chengdu, and Shanghai, which was highly relevant to their total number of residents. Compared with the official statistics from the China National Travel Administration (CNTA), the similarity of order for the top cities is much higher than that for the smaller cities such as Qingdao or Yantai. There was still a great difference between the official statistics and the social media con-

Table 1 Amount of per capita Weibo for top counties/districts

Rank	Counties/ districts	Provinces	Permanent resident population (x 10,000)	Density, per capita Weibo (item/100 capita)
1	Ejin	Inner Mongolia	2.4	9.8
2	Daocheng	Sichuan	3.1	5.0
3	Lijiang	Yunnan	42.6	3.0
4	Lixian	Sichuan	4.5	2.8
5	Jiuzhai	Sichuan	8.1	2.7
6	Changsha	Hunan	264.4	2.2
7	Sanya	Hainan	68.5	2.2
8	Kangding	Sichuan	13.0	2.2
9	Yixian	Anhui	8.1	2.1
10	Songpan	Sichuan	7.2	2.0

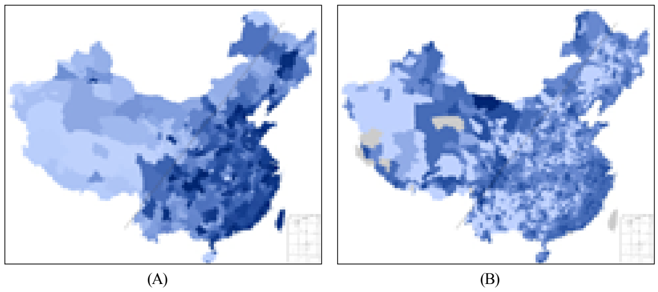

Fig. 7 Comparison between the total amount (A) and the per capita (B) of Weibo

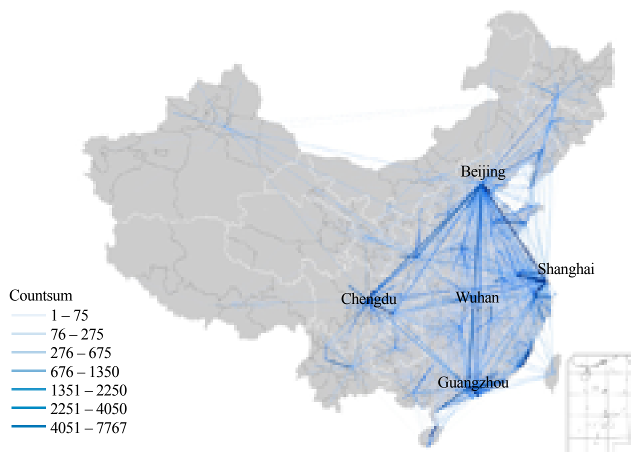

Fig. 8 Mobility network during the 2014 National Day vacation week
Source: Weibo.

sidering their sampling ways. To reduce the impact from various statistical definitions, the number of geo-located user before the vacation was used as a base to check the arrival ratio for each city. Shennongjia, Alxa, Hainan (Qinghai), and other cities got a ratio of more than 100%. These figures from on-line social platform unfolded a supplementary picture of holiday travel in China to the CNTA official statistics (see Table 2).

In addition to the major cities, several famous scenic areas were selected to investigate their tourist sources. Huangshan Mountain, Fenghuang Village, and Jiuzhai Valley got a larger diversity of tourist sources, while Wuzhen Village, Qinghai Lake, and Gulangyu Island received more local tourists. The potential tourist basins and impacts of former scenic areas were much bigger (see Table 3 & Fig. 9).

The most valuable part of social media is the content of messages themselves as well. Taking Wuzhen Village and Jiuzhai Valley as examples, by extracting the high-frequency words via the NLP

Table 2 Comparison on statistics between CNTA and Weibo

Indicators of total amount (CNTA, Weibo)					Indicators of arrivals (Weibo)		
Cities	Tourist reception (10,000)	Tourist reception (Rank)	Arrivals of Weibo users (10,000)	Arrivals of Weibo users (Rank)	Cities	Ratio	Ratio (Rank)
Chongqing	1,857.3	1	2.6	6	Shennongjia	3.37	1
Hangzhou	1,285.4	2	2.5	7	Alxa	2.69	2
Beijing	1,134.0	3	4.0	1	Hainan (Qinghai)	2.11	3
Guangzhou	1,125.5	4	3.7	2	Haibei (Qinghai)	1.93	4
Chengdu	1, 123.8	5	3.2	4	Aba	1.57	5
Shanghai	881.3	6	3.1	5	Diqing	1.39	6
Tianjin	755.5	7	1.8	13	Jiuquan	1.35	7
Suzhou	677.3	8	2.3	10	Haixi (Qinghai)	1.34	8
Yantai	371.3	9	0.7	40	Gannan	1.28	9
Qingdao	360.2	10	1.6	15	Altai	1.28	10

Source: CNTA (2014). City Rank of Profiting During the National-Day Vacation Week in 2014. <http://www.cnta.gov.cn/html/2014-10/2014-10-10-9-11-69026.html>.

Table 3 Statistics of the top 5 tourist sources of typical scenic areas

Scenic areas	Top 5 tourist sources					
	1st	2nd	3rd	4th	5th	Source regions
Wuzhen Village	Shanghai	Hangzhou	Suzhou	Nanjing	Beijing	Eastern
	34.00%	32.40%	12.80%	6.00%	3.60%	
Qinghai Lake	Huaian	Xining	Haibei	Jiuquan	Haixi	Western, central
	18.60%	18.50%	18.50%	9.10%	8.70%	
Gulangyu Island	Fuzhou	Quanzhou	Shanghai	Zhangzhou	Shenzhen	Eastern, central
	15.90%	14.20%	14.20%	11.90%	11.10%	
Huangshan Mont.	Hefei	Hangzhou	Shanghai	Nanjing	Xuancheng	Eastern, western, central
	17.40%	13.90%	13.60%	10.70%	10.50%	
Fenghuang Village	Changsha	Zhangjiajie	Wuhan	Chongqing	Huaihua	Eastern, western, southern
	38.50%	24.50%	7.30%	6.70%	5.60%	
Jiuzhai Valley	Chengdu	Chongqing	Xi’an	Beijing	Leshan	Eastern, western, central
	52.70%	15.20%	6.10%	6.00%	5.10%	

Fig. 9 Source regions of tourists of typical scenic areas

tool of Chinese languages (SCWC, 2015), a simple semantic analysis can be conducted. It was easy to find the difference between these two national-wide famous scenic areas. Most of the tourists to Wuzhen Village were from the surrounding cities, and the semantic contents were more concrete and specific, including words regarding “time” like “national day” and “one day,” or object things like “rivers and lakes,” “touring,” and “southern Yangtze river area,” and so on. In the meanwhile, the semantic contents in Jiuzhai Valley were more emotional, composed of “affection,” “beauty,” “happiness,” and so on. This type of new data made it feasible for researchers and managers of tourism to respond to the differed requests or complains for tourists from different source areas and at different periods in a near real-time manner.

4. Conclusion and discussion

Based on the previous survey and practice, this paper puts forward an analytical framework based on crowd-sourced data oriented to urban planning. Its main characters such as subjects, methods, applicable data, and so on are shown in Table 4.

Methodologies listed in Table 4 were just preliminary attempts for establishing an effective data-intensive analytical framework towards urban planning. In spite of many issues during obtaining and pre-processing the data, there are still a lot of troubles to be solved. Shan et al. (2014) summarized ten features of crowd-sourced (geographic) data as realtime, quick spread, rich with information, low cost, large amount, different quality, redundancy, heterogeneity, lack of standards, and hard to control privacy and security. All of these features demonstrate challenges and opportunities at the same time. Consequently, the following two aspects should be emphasized for further exploring crowd-sourced data in urban planning.

4.1 Integration of various methods and data sources

Using the spatio-temporal data from multiple sources respectively and synthetically, the integrated analysis will take appropriate means to handle and compare results from various types of data. Their differences and sensitivities need to be investigated. The uncertainty brought about by the uneven distribution of the crowd-sourced data should be reduced. While dealing with some “complex but sparse” data, it is necessary to introduce methods of collaborative filtering for group characteristics. In other words, deducing the missing factors for the regions with less data based on the richer ones with similar characters, can improve the reliability of results (Wang et al., 2014).

4.2 Accentuating spatial optimization for urban planning

The ultimate purpose of data analysis is to serve and improve the efficacy of urban planning and to optimize the urban entities, functional spaces, and their nexus. Spatial analysis is the prerequisite for evaluation and optimization. All of the data-intensive analysis approaches, such as statistical modeling, identifying factors, or detecting communities, should be used to promote the aim: to help recognize goals and constraints of optimization of urban planning, and further support decision-making in a quantitative way of storytelling. C

Edited by Li Caige
Proofread by Liu Jiayan

References:

Chen J, Hu B, Zuo X, and Yue Y (2014). Personal Profile Mining Based on Mobile Phone Location Data. *Geomatics and Information Science of Wuhan University*. (6), pp. 734 – 738.

Chi G, Thill J, Tong D, Shi L, and Liu Y (forthcoming). Uncovering Regional Characteristics from Mobile Phone Data: A Network Science Approach. *Regional Science*.

China National Tourism Administration (CNTA) (2014). Sichuan is the Second Largest Profiting Province During the National-Day Vacation Week in 2014. <http://www.cnta.gov.cn/html/2014-10/2014-10-9-9-4-53690.html>.

China National Tourism Administration (CNTA) (2014). City Rank of Profiting During the National-Day Vacation Week in 2014. <http://www.cnta.gov.cn/html/2014-10/2014-10-10-9-11-69026.html>.

Du Q and Ren F (2014). Representation Model of Spatial Information in Natural Language. *Geomatics and Information Science of Wuhan University*. (6), pp. 682 – 688.

Duan L, Guo W, Zhu X, and Hu B (2014). Constructing Spatio-Temporal Topic Model for Micro-blog Topic Retrieving. *Geomatics and Information Science of Wuhan University*. (2), pp. 210 – 213.

Evans J and Foster J (2011). Metaknowledge. *Science*. 331 (6018), pp. 721 – 725.

Frank L, Andresen M, and Schmid T (2004).Obesity Relationships with Community Design, Physical Activity, and Time Spent in Cars. *American Journal of Preventive Medicine*. 27 (2), pp. 87 – 96.

Gantz J and Reinsel D (2009). As the Economy Contracts, the Digital Universe Expands. <http://www.emc.com/collateral/leadership/digital-universe/2009DU_final.pdf>.

Gao X (2014). Economic and Geographic Basis for Identifying Boundaries of Cities and Urban Agglomeration by Open Data. Workshop on Big Data and

Table 4 Crowd-sourced data based analytical framework for urban planning

Subjects	Approaches		Suitable data	Comparison	Applications in planning
	Pattern	Statistics / Evaluation			
Distribution	Density (KDE)	Statistics for various units	Mobile phone positioning, weibo, etc.	Traditional statistic data; Among different periods; Among different groups of users	Identifying city center
Mobility	Clustering, detecting abnormality	Statistics for OD pairs	Mobile phone, weibo, taxi GPS, etc.		Mobility patterns of residents
Semantics	Frequent words, semantic behavior	Topic model, emotion evaluation	POIs, weibo, forum, etc.		Recognition of places
Social relationship	Social and geographic network	Betweenness centrality, node degree	Mobile phone logs, weibo, etc.		Hinterland

Spatial Development of Cities, Tsinghua University.

Gao Z and Wu J (2010). Travelers Game, Network Structure and Urban Traffic System Complexity. *Complex Systems and Complexity Science*. 7 (4), pp. 55 – 64.

Jiang B and Jia T (2011). Agent-Based Simulation of Human Movement Shaped by the Underlying Street Structure. *International Journal of Geographical Information Science*. 25 (1), pp. 51 – 64.

Li Q, Chang X, Shaw S, Yan K, Yue Y, and Chen B (2013). Characteristics of Micro-Blog Inter-City Social Interactions in China. *Journal of Shenzhen University Science and Engineering*. (5), pp. 441 – 449.

Liu Y, Liu X, Gao S, Gong L, Kang C, Zhi Y, Chi G, and Shi L (forthcoming). Social Sensing: A New Approach to Understanding Our Socio-Economic Environments. *Annals of the Association of American Geographers*.

Long Y and Liu X (2013). How Mixed is Beijing, China? A Visual Exploration of Mixed Land Use. *Environment and Planning A*. 45 (12), pp. 2797 – 2798.

Long Y and Liu X (forthcoming). Automated Identification and Characterization of Parcels (AICP) with OpenStreetMap and Points of Interest. *Environment and Planning B*.

Long Y, Gu Y, and Han H (2012). Spatiotemporal Heterogeneity of Urban Planning Implementation Effectiveness: Evidence from Five Urban Master Plans of Beijing. *Landscape and Urban Planning*. 108 (2 – 4), pp. 103 – 111.

Long Y, Zhang Y, and Cui C (2012). Identifying Commuting Pattern of Beijing Using Bus Smart Card Data. *Acta Geographica Sinica*. 67 (10), pp. 1339 – 1352.

Lu F and Zhang H (2014). Big Data and Generalized GIS. *Geomatics and Information Science of Wuhan University*. (6), pp. 645 – 654.

Lynch C (2008). Big Data: How do Your Data Grow? *Nature*. 455 (7209), pp. 28 – 29.

Niu X, Ding L, and Song X (2014). Exploring Urban Spatial Structure of Shanghai Central City Based on Mobile Phone Data. *Urban Planning Forum*. (6), pp. 61 – 67.

Sagla G and Delmelleb E (2014). Mapping Collective Human Activity in an Urban Environment Based on Mobile Phone Data. *Cartography and Geographic Information Science*. 41 (3), pp. 272 – 285.

Shan J, Qin K, Huang C, Hu X, Yu Y, Hu Q, Lin Z, Chen J, and Jia T (2014). Methods of Crowd Sourcing Geographic Data Processing and Analysis. *Geomatics and Information Science of Wuhan University*. (4), pp. 390 – 396.

Simple Chinese Word Segmentation (SCWS) (2015). <http://www.xunsearch.com/scws/>.

Sun B, Pan X, and Ning Y (2008). Analysis on Influence of Job-Housing Balance on Commute Travel in Shanghai. *Urban Planning Forum*. (1), pp. 77 – 82.

Wang J, Li C, Xiong Z, and Shan Z (2014). Survey of Data-Centric Smart City. *Journal of Computer Research and Development*. 5 (2), pp. 1 – 21.

Wu J and Di Z (2004). Complex Networks in Statistical Physics. *Progress in Physics*. 24 (1), pp. 18 – 46.

Zhang X (2014). Urban Planning Opportunity, Challenge, and Thinking in Big Data Era. *Planners*. (8), pp. 38 – 42.

Zheng Y and Xie X (2010). Intellectual Location Services Based on Mining Users' Trajectories. *Communications of the China Computer Federation*, 6 (6), pp. 23 – 30.

Zhou J, Chen X, Huang W, Yu P, and Zhang C (2013). Jobs-Housing Balance and Commute Efficiency in Cities of Central and Western China: A Case Study of Xi'an. *Acta Geographica Sinica*. 68 (10), pp. 1316 – 1330.

Zhou S and Yan X (2005). Characteristics of Jobs-Housing and Organization in Guangzhou. *Scientia Geographica Sinica*. 25 (6), pp. 664 – 670.

Zhou T, Han X, Yan X, Yang Z, Zhao Z, and Wang B (2013). Statistical Mechanics on Temporal and Spatial Activities of Human. *Journal of University of Electronic Science and Technology of China*. 42 (4), pp. 481 – 540.

Zhou X, Yue Y, Yeh G, and Wang H (2014). Uncertainty in Spatial Analysis of Dynamic Data: Identifying City Center. *Geomatics and Information Science of Wuhan University*. (6), pp. 701 – 705.

Authors

Li Dong, PhD, Senior Engineer, China Academy of Urban Planning & Design, Beijing, P. R. China. Research area: urban environment; infrastructure and spatial analysis.

Long Ying, PhD, Senior Engineer, Beijing Municipal Institute of City Planning & Design, Beijing, P. R. China. Research area: quantitative urban studies.