Perceiving Beijing’s “City Image” Across Different Groups Based on Geotagged Social Media Data

XIA PENG1,2, YI BAO3,4,5, AND ZHOU HUANG3,4,5

1 Collaborative Innovation Center of Tourism, Tourism College, Beijing Union University, Beijing 100101, China
2 State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
3 Institute of Remote Sensing and Geographical Information Systems, Peking University, Beijing 100871, China
4 Beijing Key Laboratory of Spatial Information Integration and Its Applications, Peking University, Beijing 100871, China
5 Engineering Research Center of Earth Observation and Navigation, Ministry of Education, Peking University, Beijing 100871, China

Corresponding author: Zhou Huang (huangzhou@pku.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB0503605 and Grant 2017YFE0196100, in part by the National Natural Science Foundation of China under Grant 41771425, and in part by the Beijing Philosophy and Social Science Foundation under Grant 17JDGLB002.

ABSTRACT City image in general refers to the perception, the feeling, and the opinion of a city, which contributes great importance to urban management, urban planning, urban cultural perceptions, and tourism resource development. Traditionally, city image is often inferred by the ‘five-element’ model of physical factors while lacking the consideration of subjective perception. With the rising penetration of smart mobile devices and social media, massive data of location-related texts has been generated for a variety of urban areas. The accessibility to the big data leads to a new approach of understanding the subjective perception of city image, which is important since the new approach takes the subjective heterogeneity into account. Based on the Beijing’s Weibo (microblog) data in the year of 2016, we use a random forest model to categorize user backgrounds into locals and non-locals. Meanwhile, spatial clustering is applied to identify hotspots. Then two text analysis methods—term frequency-inverse document frequency (TF-IDF) and latent Dirichlet allocation (LDA)—are adopted to abstract topics regarding the different geographical hotspots in the city across the different groups of individuals. Our research shows text mining on geotagged big data for city image makes it possible to accommodate the heterogeneity of the activities of different groups of people and to understand their preferences for different points of interests in the city, and thereby reveals the socio-cultural and functional features for the city.

INDEX TERMS City image, geotagged data, hotspots, social media, text analysis.

I. INTRODUCTION

The word ‘image’ refers to abstract feelings and perceptions, the opinions people have about a place or district, or a combination of people’s impressions, views, knowledge, judgements and sentiments [1], [2]. Capturing city image to analyze general feelings people have for different urban locations is essential to a city’s cultural inheritance and identities embodied in different urban locations. The traditional methods of perceiving city image mainly consider the physical factors of urban context, particularly represented by Lynch’s five-element model [3]. Existing research on city image can be categorized into three categories: research based on Lynch’s five-element model for element recognition in city image, research focusing on the unique elements in city image to reveal a city’s characteristics, and research that evaluates image elements to improve a city’s spatial quality [4]. However, city image differs greatly between different groups of people since they are subjectively influenced by different social status, cultural background and life experience. It is, therefore, necessary to take into account the subjective dimension when evaluating city image. A similar academic term is the geographer’s definition of ‘place’. A place is considered as a bounded space that has meaning for people [5]. It is a human-centered and semantic-enriched expression of urban location. The difference between the two concepts is that the range of city image is wider, and the image of a city can include but not be limited to its famous and iconic places.

To address the subjectivity in city image, sociological investigations have been carried out, but the methods adopted
by existing research are limited. Specifically, it is complicated to conduct surveys with large sample sizes by mental map plotting because of the analytic difficulty and the high demand for participants. Structured questionnaires also have several disadvantages such as heavy workload, small sample size, the lack of user information, and the lack of information authenticity due to subjective factors in questionnaire design [1]. For these reasons, new data sources and analysis methods for city image are required. The increased penetration of smart mobile devices and social media to people’s everyday life gives rise to the possibility.

The research objective of this work is to propose a new framework of perceiving city image by leveraging geotagged social media data, which is more efficient in data collection and more powerful in approximating subjective perception.

In recent years, people have the opportunity and platforms to express themselves via social networks anytime and anywhere on their will [6]. As location-tagging in social media smart phone applications have become mainstream, social media data contain not only users’ locations, but also the time of creation, the contents of texts, as well as photos and videos. The availability to such information results in an immense, diverse and exploitable dimension of city image. Real-time accessibility to social media allows the actual thoughts, feelings, and behavioral preferences to be immediately perceived. Therefore, anyone who posts their feelings on social media in the form of texts, photographs, emojis and tags now becomes a ‘sensor’ of multi-dimensions for the city, which is more efficient, dynamic, and diverse compared to the survey-based approach [7].

In this study, we classify Weibo users into locals and nonlocals by a random forest model based on Beijing’s Weibo geotagged data. Meanwhile, spatial clustering is applied to identify hotspots. Then, topic analysis approaches are used to extract the non-physical elements attached to the city hotspots to gain a comprehensive perception of the city image in modern Beijing. The term frequency-inverse document frequency (TF-IDF) model is used to extract the key words and the latent Dirichlet allocation (LDA) model is applied to establish a topic model to extract the two groups’ trending topics at Beijing’s hotspots. This methodology allows for a comprehensive analysis of the culture, functions and features of Beijing’s hotspots, as well as the deep mining of people’s general impressions and perceptions towards Beijing’s hotspots. The differences of the topics between the two groups are then further compared and investigated at different places in modern Beijing. On the whole, this study aims to develop a big data analytic profile for modern Beijing’s city image.

II. RELATED WORK
A. CITY IMAGE
Since 1960s, research on city image has been significantly valuable in determining a city’s identity and spatial quality, which contributes great importance to urban planning and management and also urban tourism development [8], [9]. Lynch is the first to propose an approach to city image [3]. He conducted extensive surveys and plotted mental maps to investigate citizen’s visual perception of a city, and identified five elements of city image: landmarks, nodes, paths, edges and districts. Researchers since Lynch have tended to adopt similar survey methods (i.e. questionnaires and interviews) to explore element recognition of city image. Based on Lynch’s ‘five-element’ model, Li and Yu investigated the city image of Guangzhou citizens through questionnaires [10]. Gu and Song studied Beijing’s city image using photo identification and mental maps and find the major elements in Beijing’s city image were roads, landmarks and nodes [11]. Kizil and Atalan perceived the city image of Yalova from the view of Yalova University students [12]. Kistova et al. reported results on perception of Krasnoyarsk’s city image using a decoding method [13].

However, city image is influenced by people’s subjective perception. The traditional five-element model fails to take into account the subjective dimension. Lynch’s “five-element” theory has been the basis on perception of physical elements of city image. With the rise of postmodernism, many scholars have realized that the social-economic factors, such as “social awareness, customs, history and city functions” should not be ignored in city image [4], [14]. It is also believed that city image includes not only perception of the visible, tangible elements, but also the social and cultural meanings embedded in the citizens’ activities [15]–[18].

B. GEOTAGGED SOCIAL MEDIA DATA MINING
Compared with conventional sociological methods such as mental maps and questionnaires, social media is a new way to perceive city image with low cost and high coverage. In particular, geotagged photos from social media have been popular for evaluating city image. The identification and classification of the photos’ contents can be utilized to investigate the characteristics, similarities and differences in different cities. For example, Salesses et al. used thousands of geotagged photos to compare the differences in security, social hierarchy and uniqueness between New York, Boston, Linz and Salzburg [19]. Liu et al. classified the Flickr photos using deep learning, after which they statistically analyzed the image in seven typical cities around the world to explore the relevance and diversity of city image [20]. Long and Zhou investigated the image characteristics and similarities in 24 Chinese cities through analyzing photo locations, tags and contents on Flickr [21].

Another important data source from social media is geotagged microblog. Geotagged microblog data containing texts and check-ins can directly reflect the activities, ideas and behaviors of individuals at different places. Microblog texts have been mainly used for studies on the evolution of public opinions, sentimental perception and attitude prediction towards urban issues [22]–[26]. However, there have been few studies focusing on microblog texts in regards of exploring the subjective dimension for city image.
C. TEXT MINING APPROACHES
Text mining methods are constantly evolving for topic analysis. The most typical approaches are TF-IDF method [27], [28], used frequently to extract key words, and the LDA model [29], used for topic clustering.

The TF-IDF is a commonly used weighting technique for information retrieval and data mining that uses term frequency (TF) and inverse document frequency (IDF) to evaluate the importance of a word in a document. Words with a high frequency and a low IDF are considered more important in a document and can be used as representative tags.

The TF-IDF is an explicit method for analyzing texts using visual representation; however, it is unable to capture the deep information hidden beneath the literal texts. The categorization of the words into topics can be performed using the topic clustering model. LDA is a well-known unsupervised learning algorithm that requires no manual annotation of the training sets and only requires original documents, a specific topic number k, iteration times, and Dirichlet parameters. Each document is composed of a number of words that are not specified in order, and the output of LDA is the clustering results in form of topics for these words.

D. SPATIAL CLUSTERING APPROACHES
When mining the city image from geotagged data, it is necessary to determine the spatial unit for analysis. As for big geo-data mining, the commonly used approach is spatial clustering, which could find hotspots or landmarks of the city. Through spatially clustering of the massive point sets (e.g., mobile positioning data, bus smart-card data, taxi trajectory data and social media check-in data), the boundaries of human activities can be determined, which are ‘natural’ spatial units for city image perception. Using geotagged social media data, many researchers have successively used spatial clustering to mine and identify various spatial units such as city landmarks [30], urban areas of interest [31]–[33], functional urban areas [34], attractions [35] and tourist areas [36].

However, commonly used clustering algorithms such as Kmeans and DBSCAN have some shortcomings. For example, Kmeans is unable to identify non-spherical clustering, and DBSCAN fails to achieve adaptive clustering when the spatial points are unevenly distributed. Rodriguez and Laio proposed a more robust and adaptive clustering algorithm that was able to cluster using fast search and to find the density peaks (CFSFDP) [37]. Compared with traditional spatial clustering methods such as DBSCAN, this clustering approach has significantly higher classification accuracy, is able to distinguish adjacent high-density areas and has better adaptability when there is an uneven density distribution [38].

III. STUDY AREA AND DATA DESCRIPTION
A. STUDY AREA
Spatial clustering is used in advance to extract several hotspots from the urban areas of Beijing such as Tiananmen, the Palace Museum, Jingshan and the Temple of Heaven. Each boundary of these places can be regarded as a polygon, as these hotspots (human activity-intensive areas) do not overlap and are not in contact with each other; that is, these hotspots do not cover all blocks in the Beijing urban area. With a dense urban population and high social media penetration, there are many Weibo users in Beijing, who together provide large amounts of data with distinct regional characteristics; therefore, there are sufficient sample data for perceiving Beijing’s city image.

B. DATA DESCRIPTION
Compared to other big data sources such as mobile positioning data, geotagged social media data with both texts and check-ins are considered more suitable for understanding both subjective and spatial dimensions of city image. As a check-in event is an intended action, people check-in at a place only when they are staying there for a longer period of time and believe that something is worth recording [39], [40]. Hence, we select Weibo as social media data source, since Sina Weibo is one of the most popular social media in China.

Weibo provides an open API interface for providing data. For privacy purposes, the microblog data is anonymized, but with text and check-in information. Until September 2017, Weibo had 376 million active monthly users, of which mobile users account for 92%. In this study, Weibo check-in data are used to signify the crowd activities in Beijing. Using Sina Weibo’s official open APIs, a program has been developed to capture geotagged Weibo data in downtown Beijing in 2016. The spatial heat map for the Weibo data is shown in Fig.1. There are a total of 4,654,753 check-in entries in Beijing by 1,048,575 users, each of which contains user information, post location (latitude and longitude coordinates), text content (Weibo posts) and post time. In addition, in order to classify users into locals and non-locals, we obtain the historical
check-in entries of these users by their IDs, including a total of 27,778,247 check-ins from 2013 to 2017 in both Beijing and other cities.

IV. METHODOLOGY

The overall methodological framework is shown in Fig. 2. Based on the Beijing’s Weibo (microblog) data in the year of 2016, we use a random forest model to categorize user backgrounds into locals and non-locals. Meanwhile, the spatial clustering method CFSFDP was applied to discover hotspots. We retain 73 hotspots (shown in Fig.3), which are mainly landmarks of Beijing, as the spatial units for image perception. Then two text analysis methods, i.e. TF-IDF and LDA, are adopted to abstract topics regarding the different geographical hotspots in the city across two different groups of individuals. Lastly, the city image as well as differences across the two groups are analyzed and discussed.

A. USER CLASSIFICATION

In this study, Weibo users are classified into locals and non-locals. Locals are people who have always lived or worked in the city, and non-locals are those who come to the city for a short period of time, for purposes of business or tourism. Conventional studies based on Weibo or Twitter have tended to divide users according to their sign-up locations [41], [42], but the user profiles are often incomplete or incorrect. As such classification is too simple and tends to have a high rate of false judgement, we develop a new user classification approach based on machine learning (i.e., the random forest model).

From Beijing’s Weibo data, two main differences between locals and non-locals are observed: the time interval between check-ins in Beijing and the concentration degree of check-ins outside Beijing. First, posted time intervals between the check-ins by non-locals tend to be shorter, from a couple of hours to several days; however, for local residents, the intervals could be significantly longer, some even weeks and a couple of months. Second, non-locals generally have long-term check-in actions in another city (hometown or workplace) other than Beijing, while locals mostly tweet in Beijing for a long time. Considering these differences, eigenvectors for the Weibo users are established. After manually
annoting samples, the random forest model is adopted for training. Then, a well-trained model is utilized to classify all users into locals and non-locals.

1) ESTABLISHING THE USER EIGENVECTOR
The user eigenvector mainly contains 10 features as shown in Table 1, represented by \{position_num, lat_mean, lon_mean, lat_std, lon_std, out_lat_std, out_lon_std, percentage, mean_interval, max_interval\}.

The above features are abundant for dividing users. For example, a user has a total of 50 check-in entries, four of which are geotagged in Beijing, accounting for 8%. The user’s average time interval between two adjacent check-ins in Beijing is 6 hours (rounded), and the maximum interval is one day; that is, the user stayed in Beijing for only one day and posted four check-ins. The user’s standard deviations of latitudes and longitudes outside Beijing are 0.872834 and 4.563648, which are significantly less than those within Beijing (i.e. 2.635006 and 5.196088). These features indicate that the user lives in another city rather than Beijing and is therefore more likely to be a non-local.

2) TRAINING AND PREDICTING USING THE RANDOM FOREST MODEL
After normalizing the eigenvectors, a feature matrix of all Weibo users is generated. Twenty thousand users are randomly selected for manual annotation, with ‘0’ denoting nonlocals and ‘1’ denoting locals. The 20,000 users’ eigenvectors are used as samples, and the random forest model is deployed for training and predicting.

The random forest model trains data through constructing a multitude of decision trees. It enables handling a large number of input variables, and it can assess the importance of variables when determining categories. Therefore, scikitlearn, a machine learning library for Python, is utilized to develop the RF-based classification program. The parameters for the random forest classifier are shown in Table 2. After training,
the classifier is utilized to predict types (local or non-local) for all users by inputting the feature matrix.

In order to test the classifier’s efficiency, another 20,000 users are randomly selected for manual annotation and comparison with the predicted results from the classifier. It is observed that the accuracy of our RF-based classifier is 97.7%, indicating that it enables effectively distinguish the locals and non-locals. Finally, the 1,048,575 Weibo users are divided into 796,779 locals and 251,796 non-locals.

However, our method’s disadvantage is that it is relatively complex and the training samples should be annotated manually. Another simpler idea is to classify users by multivariate statistics approaches such as regression or clustering. We plot the distribution charts of check-ins among these two groups of individuals to observe the possibility, as shown in Fig. 4. It is observed that the number of check-ins of local users accounts for the majority, and the distribution of the number of check-ins of the two groups roughly follows the long tail distribution, with no significant difference in explicit patterns. Fig. 4(A) shows the visualized 2-dimension feature space of the two groups by t-SNE (t-distributed Stochastic Neighbor Embedding), reducing the dimension of features (as shown in Table 1) of randomly selected 3,000 users. From the figure, it is observed that the two distributions are fairly mixed and coincident, and it is difficult to effectively distinguish the two groups by multivariate statistics. Although manual annotation is time-consuming, an important significance is to find and label the samples that are difficult to be distinguished by traditional statistical methods, and then learn the hidden patterns from the data through machine learning.

B. PROCESSING TEXT DATA

After 73 hotspots are identified through the spatial clustering, we reserve the check-in entries within the range of hotspots as the basis for further text analysis. The Weibo text set at each place (i.e. hotspot) is deemed to be a document, and TF-IDF and LDA methods are applied to perceive subjective image of Beijing. At the same time, because each user is identified as local or non-local, each document can be further divided into two sub-documents for locals and non-locals. Hence, both the image of the whole and the image of each group can be observed through text analysis.

The text data are processed as follows:

1) BUILD DOCUMENT SETS

All Weibo texts posted at the same hotspot constitute a long text, which is taken as a document of this place. After collecting documents from all places the document sets are built. In order to analyze the city image across the different groups, three document sets including all users, locals and non-locals are divided.

2) CLEAN THE DOCUMENTS

Through regular expression based text replacement, all documents are cleaned by removing unwanted characters such as tags, punctuation marks, emoticons, special symbols and URL links. Then, a Chinese text segmentation tool named ‘Jieba’ is utilized to segment the cleaned texts and remove the stop words.

3) TF-IDF BASED TEXT ANALYSIS

The TF-IDF algorithm is used to calculate the TF-IDF weight for each word in each document set. Therefore, the Top-10 key words at Beijing’s hotspots for all users, locals and non-locals are extracted and then compared.

4) LDA BASED TEXT ANALYSIS

The LDA algorithm is applied for topic clustering. The ‘document-word’ matrix is input and topic vectors (including the probability for each topic and the probability distribution for the words within the topic) are output. Our experiments illustrate that when the topic number N is set to 60, the perplexity of LDA outputs of Beijing’s text sets is smallest. The LDA topics at different places for all users, locals and non-locals are extracted and then compared.

V. RESULTS & DISCUSSION

A. OVERALL IMAGE

1) DISTRIBUTION OF TF-IDF TRENDING TOPICS

The TF-IDF based analyzed results can basically reflect the people’s activities and focus points at different places.

| TABLE 2. Parameters for the random forest classifier. |
|-------------------------------------------------------|
| Parameters | Description | Value |
| n_estimators | Number of trees in the forest | 100 |
| criterion | Function for measuring the quality of a split | gini |
| max_depth | Maximum depth of the tree | None |
| max_features | Number of features to consider when looking for the best split | auto |
| min_samples_split | Minimum number of samples required to split an internal node | 2 |
| bootstrap | Whether bootstrap samples are used when building trees | True |

FIGURE 4. Distribution diagrams of check-ins among locals and non-locals.
Therefore, the top-10 key words are extracted to describe the places. Table 3 lists part of the TF-IDF results. The topics that interest people at different places in Beijing can be summarized as the following four categories:

**a: LANDMARK ARCHITECTURES**

People would be attracted by its unique architectural style at various places. Time-honored, memorial, or symbolic buildings (such as the Palace Museum, Tiananmen and the Great Hall of the People), and landmark buildings of unique shape (such as the CCTV Headquarters and the Beijing National Aquatics Center) are widely mentioned by Weibo users.

**b: CITY ACTIVITIES**

Typical activities at different places are frequently discussed, such as business activities in Wangfujing, sightseeing at the Palace Museum, learning at universities and colleges, recreational activities in Sanlitun, flag raising at Tiananmen Square, watching drama at the National Theatre, visiting exhibitions at the 798 Art District, boating in Beihai Park, and large-scale concerts at the Beijing National Stadium.

**c: LOCAL CULTURE, INCLUDING HISTORY, FOLK CULTURE AND CONTEMPORARY CULTURE**

For example, at Beihai Park, the lyrics of “Let’s Paddle Together”, the theme song of the children’s film “Flowers of the Motherland”, which was produced decades ago, are frequently mentioned. At the Yonghe Temple, the key topics are burning incense and praying. At the Temple of Heaven, the most trending words are the temple fairs.

**d: FAMOUS EVENTS AND FIGURES**

For example, at Jingshan Park people often mention that the Chongzhen Emperor hanged himself here. At the Old Summer Palace, people usually mention the burning of the palace by the Anglo-French expedition forces.

In addition, the distribution of the trending term frequencies over time are also analyzed. It is observed that the topics depended on the specific features of the place and the key words are steadily fluctuating over time. In particular, topics about daily activities, folk culture and important historical events are found to be generally stable for a place.

2) DISTRIBUTION OF LDA TOPICS

Through clustering words into different topics, the LDA model is able to mine hidden information behind the explicit texts. Table 4 lists some LDA topics and the associated key words. Because LDA is an unsupervised clustering method, the semantics of the generated categories need to be identified manually. For example, Topic 9 is about going home and Topic 18 is about travelling that includes related activities such as ‘walking’, ‘sightseeing’ and ‘one-day tour’. Topic 31 is about internet start-ups, and contains some famous internet companies in Zhongguancun, Haidian district. It also covers working status of programmers, such as ‘testing’, ‘work overtime’, ‘coffee’, and ‘live-streaming’, the hottest industry trend in 2016. Topic 54 is about study, and Topic 2 is about the Old Summer Palace that explicitly includes the historical ‘burning of the Old Summer Palace’ event and the patriotic sentiment of “never forget the national humiliation”. Topic 7 is about the Beijing National Stadium, implying that the Beijing National Stadium is now used primarily as a venue for large-scale concerts. The most influential concerts in 2016 are those performed by Eason Chan and Mayday. Topic 46 is about Wangfujing, and indicates that when people are shopping around Wangfujing they are most interested in Beijing’s characteristic food and restaurants such as Quanjude, Donglaishun, Beijing Da Dong and Grandma’s Kitchen.

Through analyzing the LDA clustering results, it is found that people are interested in different topics at different places. The interests of users vary greatly even at different places of the same type. Several royal gardens, i.e., Jingshan Park, Summer Palace, Beihai Park and Old Summer Palace, are selected for comparison. The topic distribution for these places is shown in Fig.5. In the radar chart, the radius denotes topic frequency, and five typical topics are annotated with internal words sorted in descending order of occurrence frequency. It is observed that four royal gardens have topics consistent with their unique features.

The topics for the Summer Palace and Beihai Park overlap a lot, indicating that these two gardens are similar. Compared to the other two gardens, the unique features of the Summer Palace and Beihai Park are ‘lake view’ and ‘boating’, implying that people here are more willing to enjoy boating on the lake. Similar with the TF-IDF result, Beihai Park’s landmark song, “Let’s Paddle Together”, is also embodied in the LDA topic.

The closeness between Jingshan Park and the Palace Museum is fully demonstrated in the topics about Jingshan Park. As Jingshan Park has a unique function of “overlooking the entire Palace Museum”, people here tend to take more photos and often look over the central axis from Jingshan Park. At the Old Summer Palace, however, because of the national humiliation, there is a very different topic distribution from the other gardens.

When visiting a royal garden, people have diverse expressions of ‘fatigue’ (tired topic). In particular, Weibo users feel the most tired at the Summer Palace, that might be due to its large area and multiple internal attractions. At the Old Summer Palace people feel the second most tired, while at Jingshan Park and Beihai Park fewer people mention they are ‘tired’ because the two parks are relatively small.

**B. IMAGE DIFFERENCES ACROSS THE DIFFERENT GROUPS**

1) DIFFERENCES IN PLACE PREFERENCES

Comparing the spatio-temporal trajectories for locals and nonlocals in Beijing, it is observed that the two groups have different place preferences. Fig. 6 shows the heat map for the difference between locals and non-locals. The range of non-locals’ activities is relatively concentrated, mainly in the
TABLE 3. Examples of the results from the term frequency-inverse document frequency (TF-IDF) analysis.

| No. | Term 1          | Term 2          | Term 3          | Term 4          | Term 5          | Term 6          | Term 7          | Term 8          | Term 9          | Term 10         |
|-----|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| 1   | Tiananmen       | Jingshan        | The Palace Museum | Beijing National Stadium | Sanlitun        | Beihai Park     | Old Summer Palace |
| 2   | Chang’an Avenue | The Palace Museum | Forbidden City | Mayday | like | row | ruins |
| 3   | flag raising    | Forbidden City | Imperial Garden | concert | delicious | the paddles | white tower |
| 4   | Tiantanmen      | panoramic | turret | live | Magnum | boat | Garden of Gardens |
| 5   | Chongzhen       | red wall | Meridian Gate | Ashin | ten years | sea surface | reflection |
| 6   | Beijing Railway Station | Zhongguancun | school | national humiliation | East | listen | singing |
| 7   | vehicle         | testing | Weibo | classmates | stuff | never forget | broken walls |
| 8   | railway station | one-day | tour | street | discovery | garden | Eason |
| 9   | railroad station | scenery | Feiex | workhard | film | of gardens | lotus |
| 10  | next time       | sightseeing | feet | work overtime | history | Ten Years | love |
| 11  | delay           | history | building | things | black swan | one youth | lecture |
| 12  | departure       | travel | goodbye | class | royal | one teacher | teacher |
| 13  | station         | walking | coffee | morning | the garden | first time | thanks |
| 14  | hours           | building | live-streaming | future | Western House | thanks | classmate |
| 15  | subway          | wandering | Internet | days | gardens | three years | the garden |
|     |                  |                  |                  |                  |                  |                  |                  |

TABLE 4. Examples of topics from the latent Dirichlet allocation (LDA) analysis.

| No. | Topic 9 | Topic 18 | Topic 31 | Topic 54 | Topic 2 | Topic 7 | Topic 21 | Topic 46 |
|-----|---------|----------|----------|----------|---------|---------|----------|----------|
| 1   | going home | sightseeing | start-up | study | Old Summer Palace | PKU | Wangfujing |
| 2   | goodbye | visit | company | teacher | history | Beijing National Stadium | concert | Weiming Lake | roasted duck |
| 3   | train | tired | Sina | write | ruins | Mayday | Peking University | Weiming lakeside | Quanjude |
| 4   | Beijing Railway Station | place | Zhongguancun | school | national humiliation | East | Weimaingine | film |
| 5   | vehicle | tourists | testing | Weibo | classmates | stuff | never forget | broken walls |
| 6   | railway station | one-day | tour | street | discovery | garden | Eason |
| 7   | railroad station | scenery | trade | work hard | film | of gardens | lotus |
| 8   | next time | sightseeing | feet | work overtime | history | Ten Years | love |
| 9   | delay | travel | building | goodbye | class | black swan | one youth | lecture |
| 10  | departure | walking | coffee | morning | the garden | first time | thanks |
| 11  | station | building | live-streaming | future | Western House | thanks | classmate |
| 12  | hours | wandering | Internet | days | gardens | three years | the garden |
| 13  | subway | leave | around | after work | attending | class | the Eight-Power Allied Forces | good time | canteen |

city center and some famous attractions of suburbs (the red color represents the area where non-locals outnumber locals). Fig. 7 illustrates the top-10 places preferred by locals and non-locals. Except for the transport hubs (such as Beijing Railway Station, Beijing South Railway Station and the Capital International Airport), the places preferred by non-locals are mainly famous attractions such as the Palace Museum, Wangfujing, Nanluoguxiang, the Summer Palace and the Beijing National Stadium. While the locals have much lower interests in the attractions. They often visit local leisure places such as the Olympic Forest Park, the Workers’ Stadium and the 798 Art District.

2) DIFFERENCES IN PLACES TOPICS

a: NON-LOCALS ARE MORE INTERESTED IN LOCATION-RELATED TOPICS

It is observed that non-locals pay more attention to locations and related topics (e.g. for Peking University, the location-related topics are “PKU, Weiming Lake, Boya Pagoda and
FIGURE 5. Topic distribution for the royal gardens.

Weiming lakeside”). Fig. 8 illustrates the proportions of location-related topics for both locals and non-locals. At more than 65 percent of the places, non-locals are more likely to pay attention to the location-related topics than locals.

A possible explanation is that during their first visit to a place, people would pay more attention to the location-related topics. In subsequent visits to the place, people would focus more on other functions. Non-locals are more likely to visit a place for the first time, while locals are more likely to visit a place multiple times. For this reason, the topic distribution of the two groups is different. For instance, the non-locals tend to see universities as attractions and are more interested in location-related topics about the campus. In contrast, the locals (teachers or students) who live in the campus discuss more on the topics related to their lives and study. Table 5 shows the topic distribution at Peking University. For both locals and non-locals, the top-5 LDA topics are sorted in descending order of occurrence probability. It can be seen that non-locals talk most (accounting for 16.3%) about Topic 52 (i.e., the location-related topic for non-locals), while for locals the location-related topic (Topic 29) ranks only fifth (6.7%).

b: LOCALS ARE INTERESTED IN MORE TOPICS

To understand the topic concentration for Beijing’s places, the variation coefficient $\delta$ for topic distribution for each place is calculated as follows:

$$\delta = \frac{\sigma_T}{\text{Mean}(T_i)}$$

where $\sigma_T$ denotes the standard deviation of the topic probabilities, $\text{Mean}$ represents the mean value function, and $T_i$ is the occurrence probability for topic $i$. 
A higher $\delta$ value indicates that the users’ interests are more concentrated, while a lower $\delta$ value indicates that the users’ interests are more dispersed. Fig. 9 shows the variation coefficient curves for topic distributions of the two groups. It can be seen that out of 73 places, 50 (68.5%) have a higher variation coefficient for non-locals than that for locals. For instance, the locals’ $\delta$ value at Yuyuantan Park is 2.97, while for non-locals it is 3.53. Table 6 shows the top-10 topics for locals and non-locals at Yuyuantan Park. It can be seen that the locals’ topic distribution is relatively even, while the probability of the top-10 topics for non-locals goes down very quickly. As non-locals are unfamiliar with the place, they tend
to have focused interests. However, as locals are familiar with the place, they are interested in more topics.

c: NON-LOCALS ARE MORE LIKELY TO FEEL ‘TIRED’

We find that at most places (62 places) there is a “get tired sightseeing” topic. As shown in Table 7, Topic 11 is the tired topic for locals and Topic 56 is the tired topic for non-locals. Fig. 10 shows the ranking curves of the “get tired sightseeing” topic for locals and non-locals. It is observed that out of the 62 places, 55 have a higher ranking of the “get tired sightseeing” topic for non-locals than that for locals. That indicates non-locals are more likely to feel tired during their trips. One possible explanation is that as time is limited, non-locals have to travel again after a short rest. So non-locals are more likely to feel ‘tired’.

C. DISCUSSION

The result of text mining on Beijing’s Weibo data reveals the heterogeneity of the activities of different groups of people. It helps to understand people’s preferences for different points of interests in the city, and thereby reveals the socio-cultural and functional features of the city, which are meaningful for urban management and tourism resource development. For instance, from the topics of Beijing National Stadium, it is found that the stadium is as a tourist attraction and concert venue in the current usage. That indicates the

---

**FIGURE 8.** Distribution curves for the location-related topics.

**FIGURE 9.** Variation coefficient curves of topic distribution.
X. Peng et al.: Perceiving Beijing’s "City Image" Across Different Groups Based on Geotagged Social Media Data

TABLE 6. Distribution of the top 10 topics for yuyuantan park.

| Locals | Yuyuantan Park | Non-locals |
|--------|----------------|------------|
| sakura, Yuyuantan, Yuyuantan Park, Cherry Blossom Festival, spring, tours, parks, flowers | Topic 42 | Topic 11 sakura, Amitabha, the central television tower, Yuyuantan, submarine, meters, lead |
| beauty, haze, pat, sunshine, good, flowers, beautiful scenery, weekend, blue sky, blowing | Topic 9 0.17611 | Topic 0 0.27593 |
| feeling, eating, running, two, discovery, the world, only one | Topic 1 0.12752 | Topic 1 0.12968 |
| like, love, shoot, thanks, happy, work, new, lovely, time | Topic 0 0.12546 | Topic 1 0.12271 |
| love, finally, eat, hope, happy, night, life, feeling, tomorrow, friends | Topic 50 0.101 | Topic 58 like, place, love, world, pat, life, much, evening, tomorrow, photo |
| life, play, one, year, happiness, photos, children, every time, yesterday | Topic 5 0.09255 | Topic 22 park, beauty, scenery, sun, sunset, blowing, spring, lake, flowers |
| spring, scenery, blooming flowers, blooming trees, flowers, spring outing, garden, park | Topic 24 0.06985 | Topic 35 hourly, finally, the first, the end, time, the next, haze, seat, good-bye, home |
| hope, buy, end, go out, like, special, good morning, baby | Topic 2 0.05772 | Topic 2 0.07114 |
| mood, eating, things, hard work, good, no, dead, looking, face, telling | Topic 32 0.05128 | Topic 2 0.04885 |
| teachers, feeling, participation, remembrance, good, special, activity, listening, year, experience | Topic 4 0.04667 | Topic 17 art, design, new, opportunity, painting, micro-blog, appreciation, master, support, beauty |
| | | Topic 4 0.02655 |
| | | Topic 4 0.02445 |

FIGURE 10. Ranking curves of the "get tired sightseeing" topic for locals and non-locals.

TABLE 7. "Get tired sightseeing" topic.

| Topic | Words |
|-------|-------|
| Topic 11 of locals | sightseeing place visit tourists tired |
| | history one-day tour feeling scenery fatigue |
| Topic 56 of non-locals | sightseeing visit beautiful take photos tired |
| | play hot history tourists China |

stadium is not as successful as a place for public fitness and professional sports. The tag “Let’s Paddle Together” of Beihai Park illustrates that developing a cultural brand for scenic spots could be effective in attracting tourists. The differing interests of the locals and non-locals at the various places indicate how services and promotional strategies
could be adjusted for different groups of people. Specifically, locals’ experiences should be shared with the non-local tourists, which help to improve and enhance non-locals’ travel experiences. In addition, the tired topic of places indicates that more rest seats should be provided in tourism attractions.

This study not only investigates the ‘non-physical’ city image of Beijing across different groups based on text mining, but also provides a new big data based methodological framework for comprehensive perception of city image. The technical innovations are as follows. Adaptive spatial clustering is applied to process Weibo’s geotagged data to discover ‘natural’ urban hotspots. Meanwhile, a random forest model is developed to categorize user backgrounds into locals and non-locals. In particular, our machine learning based approach for user classification is much more accurate than the conventional method that divides users according to their sign-up locations. Then in order to observe the heterogeneity of the activities of different groups, two text analysis methods including TF-IDF and LDA are adopted to abstract topics regarding the different hotspots in the city.

It should also be noted that Weibo users cannot represent the total population at Beijing. The data bias needs to be overcome in the future. According to the statistics from Sina Weibo, its dominant user group is young people under 30 years old (nearby 80%), and more than half of the users have a bachelor’s degree or above. Therefore, using geotagged Weibo data might ignore some social groups, such as children, the elderly, the poor and foreign tourists. Nonetheless, compared with using the traditional social survey method such as questionnaire, we can obtain much richer information from Weibo data in a fast and low-cost way. As for Beijing’s case, the geotagged data of 1,048,575 users are used. It is quite difficult to gather information from so many users by traditional methods.

VI. CONCLUSIONS AND FUTURE WORK
In this study, Beijing’s geotagged Weibo data in 2016 is used to mine topics at different hotspots for the locals and non-locals. The city image of Beijing across two different groups of people is deeply mined through a big data based methodology comprising spatial clustering, RF-based user classification, and TF-IDF/LDA based text analysis. The methodology allows for a comprehensive analysis of the culture, functions and features of Beijing’s hotspots, as well as the deep mining of people’s general impressions and perceptions towards Beijing’s hotspots. The differences in the topics between the two groups are then further compared and investigated. On the whole, we develop a big data analytic profile for modern Beijing’s city image.

In the future, we plan to investigate more complete city image of Beijing from the big data perspective. Currently the users are only divided into two classes (locals and non-locals). We expect to propose a new classification algorithm to achieve a more sophisticated user classification (such as office workers, freelancers, young women and IT professionals). In this way, the city image across more fine-grained groups can be perceived. On the other hand, we hope to use more years of social media data for image mining, so as to explore the evolution of Beijing’s city image.

REFERENCES

[1] P. Laaksonen, M. Laaksonen, P. Borisov, and J. Halkoaho, “Measuring image of a city: A qualitative approach with case example,” Place Branding, vol. 2, no. 3, pp. 210–219, Jul. 2006.
[2] B. Kai and Z. Anzhou, “Studies on convergence and divergence of city image and destination image,” Prog. Geography, vol. 30, no. 10, pp. 1312–1320, 2001.
[3] L. Lynch, The Image of the City. Cambridge MA, USA: MIT Press, 1960.
[4] L. Q. Xu, “The rethinking of themes and paradigms: A review of urban image studies in China,” New Archit., no. 140, pp. 114–117, 2012.
[5] Y. F. Tuan, “Place: An experiential perspective,” Geograph. Rev., vol. 65, no. 2, p. 151, Apr. 1975.
[6] L. Zhou and T. Wang, “Social media: A new vehicle for city marketing in China,” Cities, vol. 37, pp. 27–32, Apr. 2014.
[7] Y. Liu, X. Liu, S. Gao, L. Gong, C. Kang, Y. Zhi, G. Chi, and L. Shi, “Social sensing: A new approach to understanding our socioeconomic environments,” Ann. Assoc. Amee. Geographers, vol. 105, no. 3, pp. 512–530, May 2015.
[8] Y. F. Tuan, Space and Place: The Perspective of Experience. Taibe: China Univ. Minnesota Press, 1997.
[9] S. Y. Zhou, S. Y. Tang, and J. C. Dai, “Identification of the significance of the concept of place to branches under human geography,” Hum. Geography, vol. 26, no. 6, pp. 10–13, 2011.
[10] Y. Li and X. Q. Xu, “A spatial analysis of the image of Guangzhou city,” Hum. Geography, vol. 8, no. 3, pp. 27–35, 1993.
[11] C. L. Gu and G. C. Song, “Urban image space and main factors in Beijing,” Acta Geographica Sinica, vol. 56, no. 1, pp. 64–74, 2001.
[12] C. Kizil and A. Atalan, “Image and reputation of Yalova City: A study on Yalova University students,” Emerging Markets J., vol. 5, no. 2, pp. 71–78, 2015.
[13] A. Kistova, N. Pimenova, and K. Reznikova, “Place management: Decoding the visual image of a siberian city,” J. Appl. Econ. Sci., vol. 11, no. 6, pp. 1143–1155, 2016.
[14] S. A. Al-ghamdi and F. Al-Harigi, “Rethinking image of the city in Saudi Arabia,” J. Siberian Federal Univ., Humanities Social Sci., vol. 8, no. 2, pp. 210–219, Jul. 2006.
[15] H. J. Klein, The Delimitation of the Town-Centre in the Image of Its Citizens, E. J. Brill, Leiden, The Netherlands: Urban Core Inner City, 1967, pp. 286–306.
[16] C. Gordon, “Human aspects of urban form: Towards a man: Environment approach to urban form and design: Amos Rapoport,” Urban Ecol., vol. 3, no. 4, pp. 385–386, 1978.
[17] O. S. Fedorova, “The analysis of cultural, architectural and artistic factors of the city image formation,” J. Siberian Federal Univ., Humanities Social Sci., vol. 8, pp. 1874–1879, Aug. 2016.
[18] P. Salieses, K. Schechtner, and C. A. Hidalgo, “The collaborative image of the city: Mapping the inequality of urban perception,” PLoS One, vol. 8, no. 7, 2013, Art. no. e68400.
[19] L. Liu, B. Zhou, J. Zhao, and B. D. Ryan, “C–IMAGE: City cognitive mapping through geo-tagged photos,” Geojournal, vol. 81, no. 6, pp. 817–861, Dec. 2016.
[20] L. Shu and Z. Zhou, “Pictorial urbanism: A new approach for human scale urban morphology study,” Planners, vol. 33, no. 2, pp. 54–60, 2017.
[21] L. Zhang and I. Pentina, “Motivations and usage patterns of weibo,” Cyberpsychology, Behav., Social Netw., vol. 15, no. 6, pp. 312–317, Jun. 2012.
[22] S. A. Golder and M. W. Macy, “Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures,” Science, vol. 333, no. 6051, pp. 1878–1881, Sep. 2011.
[23] L. Y. Wang, T. S. Zhu, B. B. Hao, “Life satisfaction among microblog users: An analysis on linguistic and behavior features,” Chin. J. Public Health, vol. 32, no. 2, pp. 225–229, 2016.
[25] C. U. I. Wong and S. S. Qi, “Tracking the evolution of a destination’s image by text-mining online reviews: The case of Macau,” Tourism Manage. Perspect., vol. 23, pp. 19–29, Jul. 2017.

[26] Q. Luo and X. Zhai, “I will never go to Hong Kong again! How secondary crisis Commun. ‘ Occupy Central’ Weibo shifted to a tourism boycott,” Tourism Manage., vol. 62, pp. 159–172, 2017.

[27] H. P. Luhn, “The automatic creation of literature abstracts,” IBM J. Res. Develop., vol. 2, no. 2, pp. 159–165, Apr. 1958.

[28] J. K. Sparck, “Exhaustivity and specificity,” J. Document., vol. 28, no. 1, pp. 11–21, 1972.

[29] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” J. Mach. Learn. Res., vol. 3, nos. 4–5, pp. 993–1022, 2003.

[30] R. Ji, X. Xie, H. Yao, and W.-Y. Ma, “Mining city landmarks from blogs by graph modeling,” in Proc. 17th ACM Int. Conf. Multimedia MM, 2009, pp. 105–114.

[31] I. Lee, G. Cai, and K. Lee, “Mining points-of-interest association rules from geo-tagged photos,” in Proc. 46th Hawaii Int. Conf. Syst. Sci., Jan. 2013, pp. 7–10.

[32] I. Lee, G. Cai, and K. Lee, “Exploration of geo-tagged photos through data mining approaches,” Expert Syst. Appl., vol. 41, no. 2, pp. 397–405, Feb. 2014.

[33] Y. Yang, Z. Gong, and L. H. U, “Identifying points of interest by self-tuning clustering,” in Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. SIGIR, 2011, pp. 883–892.

[34] J. Yuan, Y. Zheng, and X. Xie, “Discovering regions of different functions in a city using human mobility and POIs,” in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining KDD, 2012, pp. 186–194.

[35] X. Zhou, C. Xu, and B. Kimmons, “Detecting tourism destinations using scalable geospatial analysis based on cloud computing platform,” Comput., Environ. Urban Syst., vol. 54, pp. 144–153, Nov. 2015.

[36] H. Q. Vu, G. Li, R. Law, and B. H. Ye, “Exploring the travel behaviors of inbound tourists to hong kong using geo-tagged photos,” Tourism Manage., vol. 46, pp. 222–232, Feb. 2015.

[37] A. Rodriguez and A. Laio, “Clustering by fast search and find of density peaks,” Science, vol. 344, no. 6191, pp. 1492–1496, Jun. 2014.

[38] X. Peng and Z. Huang, “A novel popular tourist attraction discovering approach based on Geo-tagged social media big data,” ISPRS Int. J. Geo-Inf., vol. 6, no. 7, p. 216, 2017.

[39] A. M. Kaplan and M. Haenlein, “Users of the world, unite! The challenges and opportunities of social media,” Bus. Horizons, vol. 53, no. 1, pp. 59–68, Jan. 2010.

[40] J. Cai, B. Huang, and Y. Song, “Using multi-source geospatial big data to identify the structure of polycentric cities,” Remote Sens. Environ., vol. 202, pp. 210–221, Dec. 2017.

[41] Z. Gu, Y. Zhang, Y. Chen, and X. Chang, “Analysis of attraction features of tourism destinations in a mega-city based on check-in data mining—A case study of shenzhen, China,” ISPRS Int. J. Geo-Inf., vol. 5, no. 11, p. 210, 2016.

[42] H. Shao, Y. Zhang, and W. Li, “Extraction and analysis of city’s tourism districts based on social media data,” Comput., Environ. Urban Syst., vol. 65, pp. 66–78, Sep. 2017.