Chin. Geogra. Sci. 2021 Vol. 31 No. 1 pp.14–26
https://doi.org/10.1007/s11769-020-1172-6

Springer Science Press
www.springerlink.com/content/1002-0063

Analyzing Asymmetric City Connectivity by Toponym on Social Media in China

YE Xinyue[1], GONG Junfang[2, 3], LI Shengwen[2]

(1. *Department of Landscape Architecture and Urban Planning Texas A&M University*, *College Station*, *TX* 77843 *USA*; 2. *School of Geography and Information Engineering*, *China University of Geosciences*, *Wuhan* 430074, *China*; 3. *Key Laboratory of Urban Land Resources Monitoring and Simulation*, *Ministry of Natural Resources*, *Shenzhen* 518034, *China*)

Abstract: The connectedness between cities has become one of the most widely discussed topics in urban and regional research in the mobile and big data era. One problem identified is the asymmetric city connectivity, partially due to data availability. We present a data-driven approach based on location and toponym (place name) extracted from social media data, to assess the asymmetric connectivity between cities. The assumption is that a higher frequency of occurrences of the name of city *i* in posts located in city *j* would imply that the city *i* is more influential than other cities upon city *j*. In addition, we've developed a group of measurements such as the relatedness index, impact index, link strength index, dependence index, and structure similar index to characterize such interactions. This framework of connectivity measurements can also be used to support smart planning taking into account the evolving interplay among cities. The space-time structure of urban systems in China is examined as the case study.

Keywords: asymmetric city connectivity; social media; urban system; China

Citation: YE Xinyue, GONG Junfang, LI shengwen, 2021. Analyzing Asymmetric City Connectivity by Toponym on Social Media in China. *Chinese Geographical Science*, 31(1): 14–26 . https://doi.org/10.1007/s11769-020-1172-6

1 Introduction

The availability of geotagged social networking data has become a goldmine for urban and regional analysts (Zhen et al., 2019). Since social networking platforms allow the users to submit the locations from where they post, these posts may contain different types of geographic information constituted by the location of the user's profile or the location where the post is delivered (Jiang et al., 2018). Usually, very few posts are geo-tagged, so only a limited number of posts are available with spatial location (Ghosh and Guha, 2013). Other locational information may be detected by geocoding the posts' IP address (Tsou et al., 2013). At the same time, another important element of the information included in social media messages is the name of the place. A great amount of geographic information is embodied in the text (Liu et al. 2014a; Li et al., 2017; Hu et al., 2017). Many efforts have thus been made to mine geographical knowledge from text documents (Buyukkokten et al., 1999; Borges et al., 2011; Scellato et al., 2011). Methods have also been developed to improve toponym disambiguation (Zhang and Gelernter, 2014; Ghufran et al., 2015).

Human activities take place within a geographical context, so it is not surprising that many human-generated

Received date: 2019-12-25; accepted date: 2020-04-10
Foundation item: Under the auspices of National Natural Science Foundation of China (No. 41801378, 42071382), the Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources (No. KF-2019-04-033)
Corresponding author: LI Shengwen. E-mail: swli@cug.edu.cn
© Science Press, Northeast Institute of Geography and Agroecology, CAS and Springer-Verlag GmbH Germany, part of Springer Nature 2021

online contents have a location element. It has been estimated that 13%–15% of web queries contain a place name in search engines (Sanderson and Kohler, 2004). Through toponyms, such requests can be associated with the geo-spatial world (Hecht and Moxley, 2009). In recent years, much attention has been paid to the process of retrieving geographic information from unstructured online documents (Jones and Purves, 2008). Density surface modeling has also been employed to examine ambiguous places (Jones et al., 2008).

It is estimated that about 10% of tweets contain toponym information (Kitamoto and Sagara, 2012). A temporal investigation of place names also helps the understanding of socioeconomic transitions (Luo et al., 2010; Wang et al., 2012). Co-occurrences of toponyms on the web and social media have been used for place name disambiguation (Overell and Rüger, 2008; Wang et al., 2010). Wu et al. (2020) combined toponym co-occurrence with improved field strength model to identify the urban hinterland in the Pearl River Delta in China. Tsou et al. (2013) mapped social activities and concepts by integrating Twitter and web search outcomes. Spatial components on social media have also been widely utilized in public health (Hay et al., 2013; King et al., 2013; Padmanabhan et al., 2014; Velardi et al., 2014; Widener and Li, 2014; Grajales III et al., 2014), event detection (Sakaki et al., 2013), human behavior (Liu et al., 2014b), spatial planning (Massa and Campagna, 2014), and urban structure studies (Gao et al., 2013). Toponyms in social media can shed light on people's food preference elicited by exposure to their immediate food environment (Chen and Yang, 2014). In addition, toponyms have been geotagged for reading news and scientific data streams with maps (Samet et al., 2014).

Cities grow and decline in an interdependent system of flows, connections, and relations. As Beaverstock et al. (2000) stated, 'the world-cities literature is seriously unbalanced: it has a surfeit of interesting theoretical concepts for treating the nodes of the world-city network, but these exist alongside a deficit in empirical concern for measuring relations between the nodes'. City connectivity can be explored using the large-scale tracking data of flights and Google Hyperlinks (Jiang and Jia, 2011). Toponyms in the big data environment also provide new data sources for assessing city connectivity, prompting the idea that the world should be viewed as networks rather than the mosaic of cities (Beaverstock et al., 2000). Toponyms have been grouped and assessed based on how geographic factors differ among them (Wang et al., 2014). Liu et al. (2014a) proposed a method for capturing 'relatedness between geographical entities' based on the co-occurrences of their names on web pages. The relatedness measured by the co-occurrences of the cities' names (toponyms) on large web pages can be deemed as the weight of intercity urban flows (Lin and Li, 2015). A global awareness map based on toponyms was developed in one such study for cities in the United States through tweets (Han et al., 2015).

Each city can be considered as an areal unit because it is composed of certain districts, but can also be treated as a node in the urban system. The difference among these cities is not only due to the geographical size of the area covered, but also other attributes such as economic activities. Such differences would render an asymmetric relationship among cities. It is the aim of this study to measure such asymmetry. In other words, the degree of concerns of city *i* residents on city *j* (such as online discussions) is different from the degree of concerns of city *j* residents on city *i*. Cities that are more concerned tend to be more influential or dominating in the urban system. Hence, the quantification of such asymmetric relationship can provide proof for regional and urban planning.

The co-occurrences of toponyms cannot distinguish the asymmetric connectivity between cities because such a method only considers the number of times two cities co-occur without differentiating their respective weights in this relationship. Furthermore, relations between cities are dynamic. The change in transportation systems, such as the construction of a railway or highway between the two cities, will quantitatively or even qualitatively change the relationship between them.

In this paper, we highlighted the importance of asymmetry of urban connections. The major contributions of the study include: 1) The development of a methodological framework that examines city relationships taking into account the asymmetry. 2) The case study of China, using data from Sina Microblogging, shows the asymmetry of relatedness between cities is significant, and that the link strength is correlated with the joint factors of population and GDP.

2 Methodology

This research proposes a new method to explore the connectivity between geographical entities based on user location information and toponyms from posts collected from microblogs. Sina Microblog is a Chinese microblogging (Weibo) website supported by mobile apps with a market share similar to Twitter in the USA (Ye et al., 2018). Akin to a hybrid of Facebook and Twitter, Sina Microblog was launched by SINA Corporation on August 14, 2009, and remains one of the most popular social media services in China. A web crawler was used to retrieve posts with toponyms. Generally speaking, the toponyms in a post imply that the geographical entities (such as places *i*, *k*) are referenced by the place where the message is posted (*j*), and the number of posts indicates the strength of relationship between these two places *i* and *j*, *k* and *j*. Because of the rich and timely information in social media content, the toponym mentioned and the place where the message was posted can reveal the dynamic relationship between these two places. In addition, the asymmetricity connectivity can reflect the uneven flow between cities due to the different status of cities in the urban system.

Toponym occurrences in a post reveal the following four scenarios: 1) Interaction, e.g., 'After smoggy Beijing enjoyed a rare sight, you sent your rainbow pics from around the world'; 2) Status, e.g., 'Beijing, I am here'; 3) Event, e.g., 'Beijing is hosting 2022 Winter Olympics'; and, 4) News, e.g., 'Chinese #MH370 relatives stage an emotional protest outside the Malaysian embassy in Beijing'.

The first two scenarios reflect that the geographical entity, 'Beijing', is related to the location of the posts by users. The latter two scenarios play an important role in communication, even as they reveal the influence of the 'geographical entities' among users. In fact, The First Law (TFL) of geography indicates that nearer things are more related than distant things (Tobler, 1970). The term relatedness can be interpreted as spatial interaction represented by flows of individuals or attention between two localities (Miller, 2004).

The proposed method provides both an active and passive pattern to estimate asymmetric relationships between places based on the location and toponym in a post's text on social media. The urban system of China is explored using this method. We also develop a group of measurements such as the asymmetric relatedness index, link strength index, impact index, dependence index, and structure similar index based on the bidirectional relatedness between geographical entities. This paper provides a bidirectional perspective to estimate the asymmetric relationship of places, which is an extension of the work by Liu et al. (2014a) and Hu et al. (2017), which are based on the symmetric relationship of places. Additionally, this paper provides a group of new statistics for spatial impact, dependence, and similarity based on the asymmetric relatedness. With the benefit of time-stamped posts in social media, temporal relatedness can also be analyzed. The measurements of asymmetric relatedness, relatedness strength, impact, dependence, and similarity are shown in Table 1.

2.1 Asymmetric relatedness index (A index)

Let P_{ij} denotes the number of posts located in city *i* and mention city *j* in the post contents. It represents the references to the number of posts regarding city *j* by city *i*. Similarly, P_{ji} denotes the references to the number of city *i* by city *j*. Both P_{ij} and P_{ji} contribute to the strength of relatedness between city *i* and city *j*. However, these two indicators very possibly have different values. In other words, the relatedness between these two cities is bidirectional. Due to the long tail distribution (Liu et al., 2014a), we conduct a transformation as follows:

Table 1 A framework of connectivity measurements

Index	Index name	Explanation
A	Asymmetric relatedness index	A_i denotes asymmetric relatedness index of city *i*
L	Link strength index	L_{ij} denotes the strength of link betwen city *i* and *j*
I	Impact index	I_i denotes impact index of city *i*
D	Dependence index	D_i denotes dependence index of city *i*
S	Structural similarity index	S_{ij} denotes the structural similarity betwen city *i* and *j*

$$p_{ij} = \log P_{ij} \tag{1}$$

When a city is treated as a node and the relatedness between cities can be considered as a link between nodes, we can compute the geographical entities and relatedness among them as a social network system (Wasserman and Faust, 1994). According to social network theory, the out-degree and in-degree are important indices of a node. After standardizing the out-degree and in-degree of a City, we denote A_i as the asymmetric relatedness index for city i:

$$A_i = \sum_j^n \frac{p_{ji} - p_{ij}}{p_{ij} + p_{ji}} \ (k \neq j). \tag{2}$$

A_i ranges from –1 to 1. $A_i < 0$ means city i pays more attention to other cities than is paid attention to.

2.2 Link strength index (L index)

The gravity model is widely used to measure the link strength between cities (Erlander, 1980; Haynes and Fotheringham 1984; Porojan, 2001). G_{ij} denotes the relatedness between city i and city j:

$$G_{ij} = \frac{X_i X_j}{W_{ij}^{\beta}}, \tag{3}$$

where X_i and X_j are the size of two geographical entities i and j (i.e., population or GDP), W_{ij} is the spatial distance between them, β is used to adjust the impact of distance, and G_{ij} usually represents the correlated index between geographical entity i and geographical entity j. However, the gravity model is an indirect method and the value of G_{ij} is highly dependent the value of β.

In the study, the link strength between two geographical entities can be measured using the standardized value of the reference number between them. Thus, C_{ij} is denoted as the one-way link strength index between city i and city j:

$$C_{ij} = \frac{p_{ij}}{\sum_k^n p_{ik}} (k \neq j) \tag{4}$$

where C_{ij} ranges between 0–1, and its expected value is:

$$E(C_{ij}) = \frac{1}{n-1} \tag{5}$$

We can use the harmonic mean, L_{ij}, to denote the strength of the link between them:

$$L_{ij} = \frac{2C_{ij}C_{ji}}{C_{ij} + C_{ji}} (i \neq j) \tag{6}$$

2.3 Impact index (I index) and dependence index (D index)

The impacts of a city received and brought are asymmetrical. Two indices, the impact index and the dependence index are defined to measure the impacts between a focal city and other cities. The impact index measures the capacity the focal city affecting other cities. The dependence index (D index) is defined to examine the impacts received.

The impact index, I_i denotes the impact of city i:

$$I_i = \frac{\sum_j^n C_{ji}}{n-1} (i \neq j) \tag{7}$$

The values of I_i ranges between 0–1, and its expected value is:

$$E(I_i) = \frac{1}{n-1} \tag{8}$$

Due to the differences across cities regarding their attributes in population density, internet access, and infrastructure level, there exists be a great variation of interdependence between city i and city j. Therefore, the self-relationship of one city is adopted as a standard to measure the dependence from one city to other cities. The dr_{ij} variable indicates the dependence index from city i to city j:

$$dr_{ij} = \begin{cases} \dfrac{p_{ij}}{p_{ii}}, \ p_{ij} < p_{ii} \\ 1 \end{cases} \tag{9}$$

A city is expected to be more related to itself than other cities. In other words, p_{ij} should be less than p_{ii}. Therefore, dr_{ij} should range between 0 and 1. Similarly, the ratio for city i, of its out-degree and its in-degree indicates the dependence of city i. We use dd_{ij} to denote the dependence of city i as well.

$$dd_{ij} = \frac{p_{ij}}{\sum_k^n p_{ik}} (k \neq j) \tag{10}$$

d_{ij} demonstrates the comprehensive measure of external dependence between city i and city j:

$$d_{ij} = \sqrt{dr_{ij} dd_{ij}} = \sqrt{\frac{p_{ij}^2}{p_{ii} \sum_j^n p_{ki}}} (i \neq j) \tag{11}$$

The dr_{ij}, dd_{ij}, d_{ij} dependence can also be considered an asymmetric dependence relying on the ratio of city i to city j on their features. The asymmetric dependence is very difficult to estimate using traditional methods. The three indices demonstrate relationships and the difference in dependencies between cities. D_i shows the dependence of city i using a summarization of dependence to other cities:

$$D_i = \sum_j^n d_{ji} (i \neq j) \tag{12}$$

D_i is the index of dependence of a city to other cities, which is a type of relatedness of dependence.

2.4 Structural similarity index (S index)

Structural similarity indicates the degree of similarity between the two cities in the correlation based on impact and dependence. We use S_{ij}, the Pearson's correlation coefficient, between the two vectors associated with i and j, to measure the similarity:

$$S_{ij} = \frac{(n-1)\sum_k^n (C_{ik} - m_i)(C_{jk} - m_j)}{\sqrt{\sum_k^n (C_{ik} - m_i)^2} \sqrt{\sum_k^n (C_{jk} - m_j)^2}} (k \neq i, j) \tag{13}$$

$$m_i = \frac{\sum_j^n C_{ij}}{n} (i \neq j) \tag{14}$$

3 Results

3.1 Data collection

China has 32 provincial units and two Special Administrative Regions (Hong Kong and Macao). The spatial distribution of the 32 provincial capital cities is relatively dispersed and clustered in the eastern area, because the economy is relatively more developed in the eastern part of China. There are large percentages of people who have access to social media in the eastern provinces. On the other side of China, the size of cities and its population is more varied. The 50 capital cities in the USA are more diverse compared to the 32 provincial cities in PRC. Furthermore, the past 10 years have seen rapid development in the economy and technology in PRC which has produced a different relatedness among those geographical entities over time. Taiwan, Hong Kong and Macao were removed from the list of cities in this experiment because most users in the three cities use traditional Chinese, thus the data obtained here may not accurately reflect their relationships with other cities. Therefore, we chose the 31 provincial capital cities for this study.

Fig. 1 shows the map of spatial distribution of the 31 cities. This map shows that the distribution of Chinese cities is uneven with clusters of cities concentrated in the east.

Sina Microblog provides a search API (Application Programming Interface) based on location and keywords, so users can get the number of reference counts between geographical entities based on location, toponym and timespan. There is a series of search pages on the Sina Microblog website which can be used as a tool to observe time, region (province or city), and keywords. The search page not only displays the first 50 microblog posts that meet the criteria but also shows the number of posts that match the search criteria at the bottom of the page. We programed a web crawler to simulate the page search process. After running the crawler, we obtained a reference matrix of 31 cities from the posts originating in the frame of one month from January 1, 2015 to January 31, 2015.

The maximum of the reference numbers is 618 595, the minimum value is 0, and the average value and the standard deviation are 7480 and 54 353.53. Fig. 2 shows the reference numbers distribution of occurrences on logarithmic scales. The top 10 reference numbers are large than 100 000, and the lowest 900 reference numbers are less than 1000, which clearly exhibits a heavy-tailed distribution. This is consistent with the findings of similar research that use toponyms of co-occurrences on web pages (Liu et al., 2014a).

Fig. 1 The spatial distribution of the 32 provincial capital cities (Downloaded from http://bzdt.ch.mnr.gov.cn/)

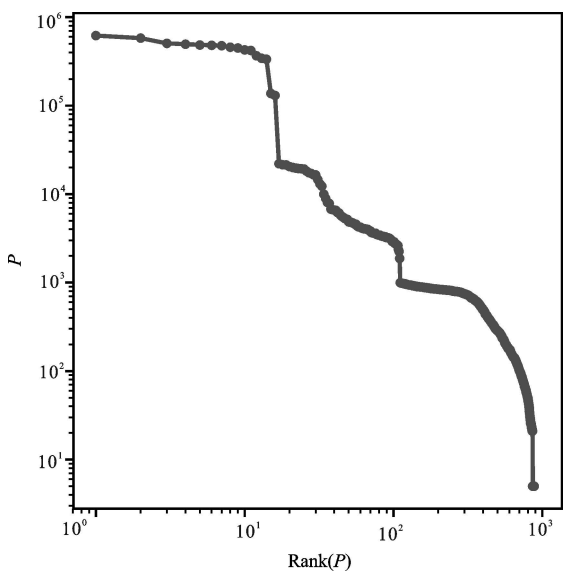

Fig. 2 Heavy-tailed distribution of toponym occurrence. *P* denotes the reference numbers

3.2 Asymmetric bidirectional relationship

In section 2.1, p_{ij} denotes the reference level from city *i* to city *j*. Although the values of p_{ij} and p_{ij} are different when *i* and *j* are not same, both them represent the relationships between two cities. In this section, we examine the asymmetric city relationship within the asymmetric relatedness index (A index) as a whole. The Pearson correlation analysis is also used to measure the correlation between the values of p_{ij} and p_{ji}. The results show that a Pearson correlation coefficient (*R*-value) is 0.063, and *P*-value is 0.734. It indicates that the linear correlation between the A index and *G* is very weak and statistically insignificant. It also means that there is a great difference between the attention and the concern indices (–0.87952, 0.83071, –0.0001841, 0.40301246). The asymmetric relatedness index of 31 cities is calculated and plotted in Fig. 3. As shown in Fig. 3. the two-way relationships between cities is very different.

Three measurements *dr*, *dd* and *d* are used to describe the dependency. Similarly, Beijing is chosen as a case to examine the asymmetric relationships between the cities in Fig. 4.

In Fig. 4(a), most cities seem to have a slightly high dependency index to Beijing, while Beijing lacks dependency on other cities. As the capital of the country, Beijing is the political center and the second largest city, in addition to being a vibrant city. Younger users from the Sina microblogging find it to be a city of, more opportunities and visions. In Fig. 4(b), we normalize the index with the total links between cities. The results show that Beijing has a similar dependency index when compared to other cities. In Fig. 4(c), the dependency relatedness of Beijing is similar to that in Fig. 4(b). In

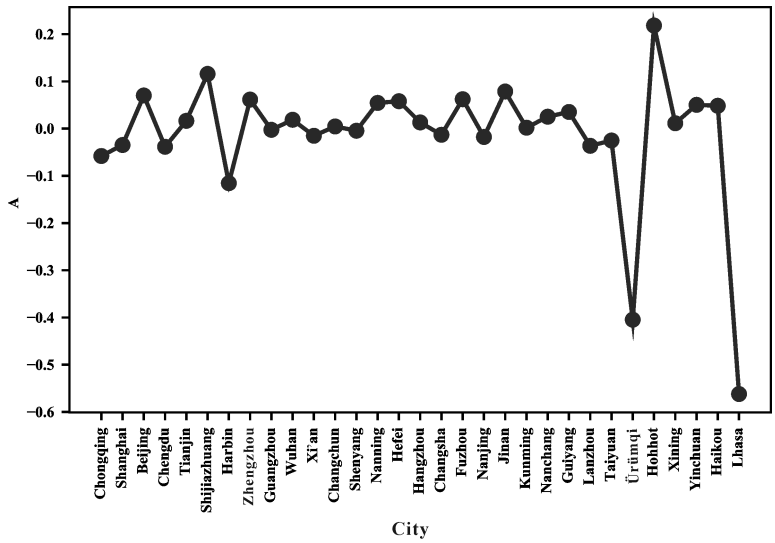

Fig. 3 The values of Asymmetric relatedness index (A) of 31 cities in China (not including Taiwan, Hong Kong and Macao)

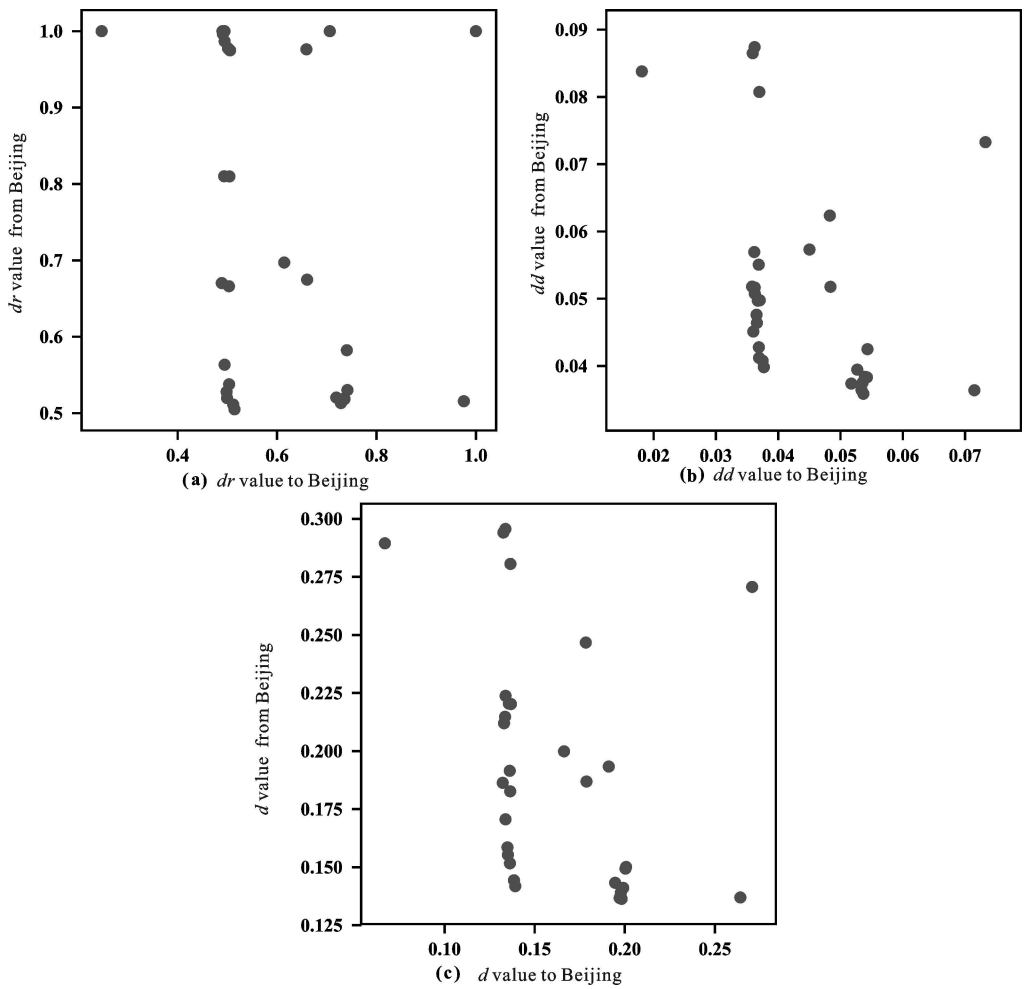

Fig. 4 Dependency related to Beijing in three measurements, (a) *dr*, (b) *dd* and (c) *d* of all cities

summary, the strength of the link is asymmetrical between cities. It is possible to measure the asymmetrical relatedness between geographical entities based on the number of references in social media.

3.3 Link strength index and the distance decay effect

At first, we designate W_{ij} as the physical distance (km) between cities as W, and set β to 1. Then the values of G, the gravity model, is calculated, and are labelled as group 'Population', and are based on the population of the geographical entities. After that, we recalculate G with GDP instead of the population and get the data tagged 'GDP'. At the same time, we also employ population and GDP at the same time as the joint factor because the activity level of social media usage should be related to that, which is noted as POP-GDP.

$$D_{ij}^{POP-GDP} = \sqrt{Population_i \times GDP_i \times Population_j \times GDP_j} \tag{16}$$

where $Population_i$ is population of city i, and GDP_i is the GDP value of city i, we get the link strength index by measuring the values of L. Lastly, we take the maximum value of each group as 1, and standardize the three group results. The Beijing example is shown as Fig. 5.

Fig. 5 shows the following: 1) Shanghai, Zhengzhou, Hangzhou, and Nanjing have demonstrated a high degree of strength in group 'Population' and 'GDP'. On the other hand, since Lhasa and Ürümqi are far away from Beijing, they have a weaker relatedness in group population and GDP. 2) The R-curve shows that relatedness measured by social media is influenced less in terms of population by the distance when compared to the traditional gravity model. However, the joint factor of population and GDP shows the statistical significance at the 0.05 level. In addition, the R-value in Ürümqi has an exception due to network access restrictions. It shows that there are still certain restrictions to measure relatedness between cities based on reference counts in social media.

3.4 Impact index and dependence index

We estimate the impact index (the I index) and dependence index (the D index) of 31 cities, and then provide a comparative analysis with their GDP and population, as shown in Fig. 6.

The I index has a curve with GDP and population (tagged as 'POP-GDP') that is similar to Fig. 6. The Pearson 2-tailed correlation is calculated to detect the relationship between the I index and POP-GDP. The value between them has a higher correlation with a value of 0.631 and P = 0.001, which is statistically significant.

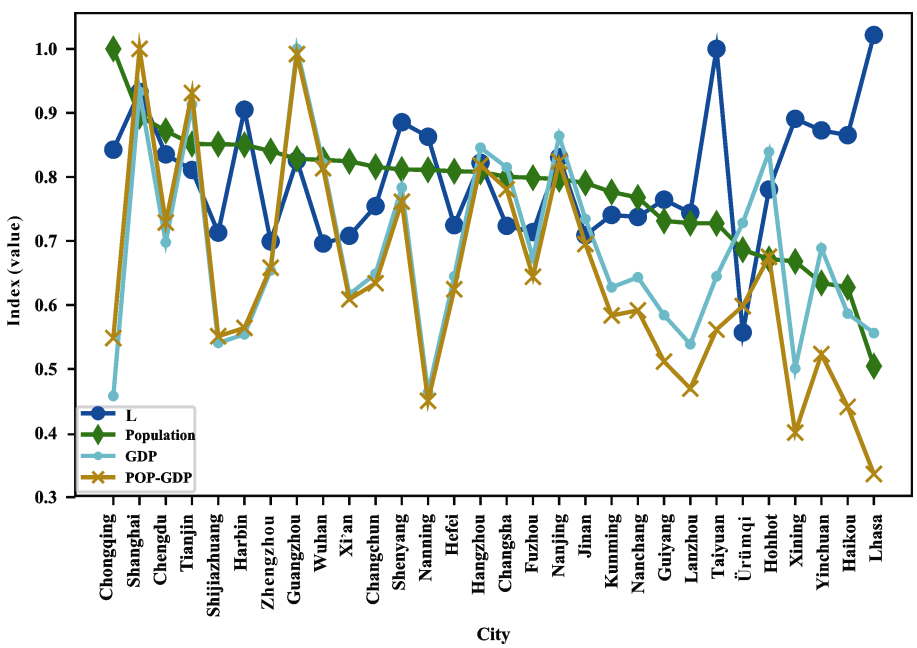

Fig. 5 Gravity model and connective strength to Beijing

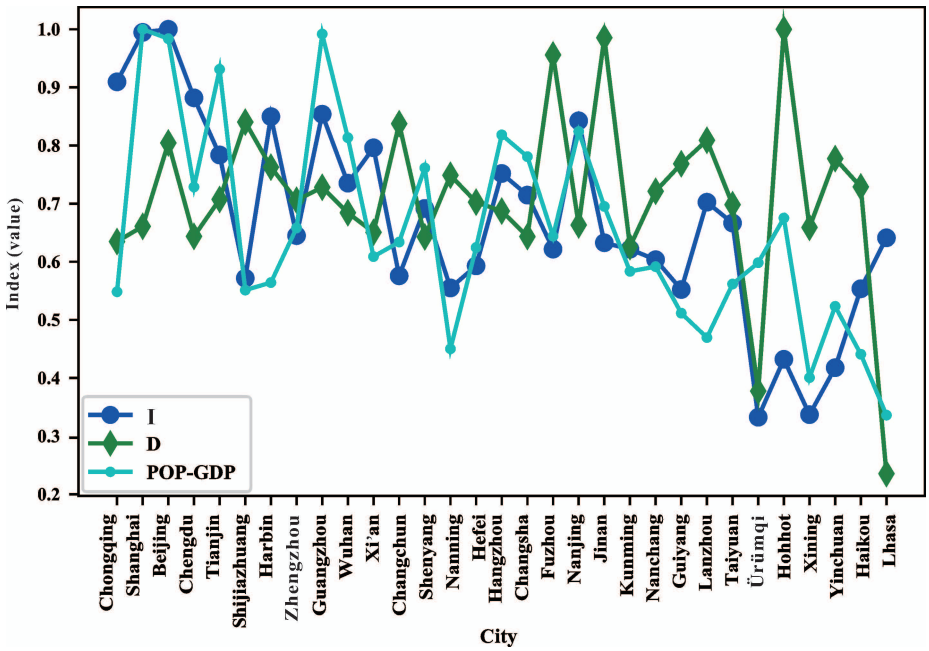

Fig. 6 Comparison Between the I (the impact index), D (dependence index) and POP-GDP

In short, the I index demonstrates a high correlation with both POP-GDP. This means that not only can *I* be measured by the relationship between the cities, but the I index is also the new comprehensive index of a city.

We also examine the number of toponyms mentioned by different cities. The results show that the number of toponyms mentioned by other cities is 97.8 percent lower than the number of times the city itself is mentioned. This means that social media has a high auto-correlated characteristics. It can be seen in Fig. 6 that 1) there is a great difference between the I index and D index ; 2) the D index has a slightly reversed trend with the I index. This result is consistent with expectations.

3.5 Structural similarity

Beijing, the political and cultural center of China, and Shanghai, the economic center, were both chosen as examples, for an observation of the similarities between them and other cities. Their similarities in comparison with other cities are shown in Fig. 7.

Fig. 7 shows that: 1) As first-tier cities, Beijing and Shanghai, which are political, cultural, and economic centers respectively, were very close in their similarities with each other, as well as to other cities. To Beijing, Shanghai is the most similar city. To Shanghai, Beijing is the most similar city too. 2) Although there is a significant difference between Beijing and Shanghai in the role and location, they both bear a similar relatedness to other cities.

There are three major metropolitan regions in China. They are the Jingjinji Metropolitan Region which includes cities such as Beijing, Tianjin and Shijiazhuang; the Yangtze River Delta Metropolitan Region which includes cities such as Shanghai, Nanjing and Hangzhou; and the Pearl River Delta Metropolitan Region which includes cities such as Guangzhou and Shenzhen. We use six cities from the first two Metropolitan Regions to observe the similarities shown in Table 2 and Table 3.

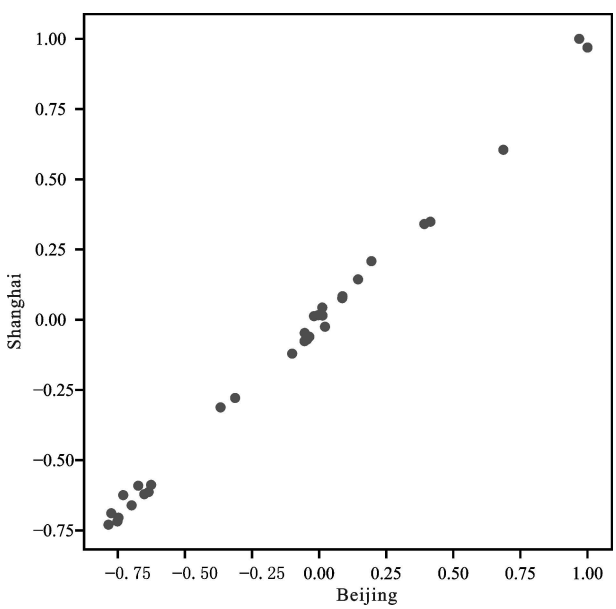

Fig. 7 Beijing and Shanghai's similarities to other cities

Table 2 Structural similarity within the Jingjinji Metropolitan Region

City	Beijing	Tianjin	Shijiazhuang
Beijing	1.000	–0.608	–0.005
Tianjin	–0.747	1.000	–0.105
Shijiazhuang	0.013	–0.142	1.000

Table 3 Structural similarity within the Yangtze River Delta Metropolitan Region

City	Shanghai	Nanjing	Hangzhou
Shanghai	1.000	0.007	–0.573
Nanjing	–0.231	1.000	0.018
Hangzhou	0.013	–0.023	1.000

Table 2 shows that: 1) Cities in the same Metropolitan Region are not similar. Instead, there appears to be a possible complementary relationship; 2) Beijing and Shanghai have fewer similarities with other cities in their Metropolitan Regions. 3) Beijing and Shanghai, Tianjin and Nanjing, Hangzhou and Shijiazhuang, are three pairs of similar cities within these two Metropolitan Regions. Each group plays a similar role in in their Metropolitan Regions. Hence, such similarities could be used to identify the role and relationship among cities in Metropolitan Regions.

Based on the value of similarity S_{ij}, the nearest neighbor classification method is used to classify the 31 cities as shown in Fig. 8. Out of the 31 cities listed, 15 of them formed 6 initial clusters. The sixth cluster is made up of Wuhan, Hangzhou and Changsha, which have two relatively close cities, Wuhan and Changsha in the distance. The other 5 clusters are not located in adjacent provinces. It can be concluded that the similarity between the cities is partially related to geographical distance. In short, the S index provides a new method, with high stability and consistency, to measure inter-city similarity.

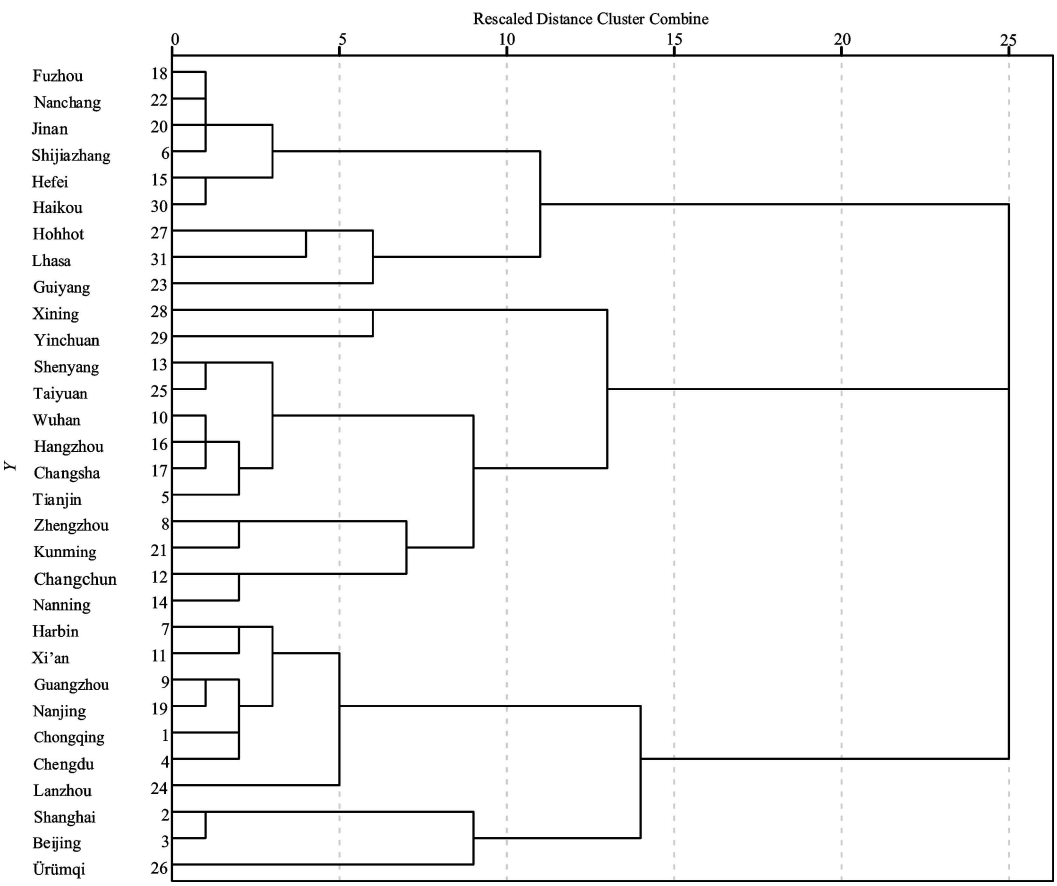

Fig. 8 Cluster analysis based on the similarity of 31 cities in China

3.6 Temporal factor

Out of the 31 cities, Beijing, Shanghai, Tianjin and Chongqing are the only four provincial cities. They are not geographically clustered and play different roles in China. Beijing is the political and cultural center of China, Shanghai is the economic center of China, Tianjin is a fast-growing economic center and Chongqing is one of the four municipalities (the other three are Beijing, Shanghai and Tianjin) under the direct administration of central government and the only one located in the inland. Geographically, Beijing and Tianjin are close. Both of these cities are located in North China. Shanghai is located in the eastern China. The four cities were chosen as examples to observe regional impact changes over time. We measured the impart index, *I*, of the four cities on a monthly basis from 2011–2015.

Fig. 9 shows that 1) at the beginning, Beijing and Shanghai had a higher urban influence. The impact gradually decreased after that, and then rose in 2014. 2) Compared to Beijing and Shanghai, the impact of Chongqing and Tianjin had a reverse trend. Their impact was low in the beginning, and then gradually increased before 2014 and then fell in 2014.

In short, the impact of the four cities is becoming more and more unified. Since the data on social media has natural spatial and real-time characteristics, it can be applied to measure the relatedness between cities in different time scales and time intervals. So, the relatedness estimation based on social media is not only a multi-scale measure method, but also supports the ability to predict the trends of relatedness among cities.

4 Conclusions

Human dynamics reflected in big social media data offer grand opportunities to measure and compare the flows of information between cities. Based on the spatial information retrieved from social media, including toponyms and the location of posts, we propose an effective method to estimate the relatedness, impact, and similarity of cities. Both symmetric and asymmetric relatedness have played an important role in measuring and comparing city relationships. The proposed method using social media data provides both an active and passive pattern to estimate the asymmetric relationship between cities. Additionally, the suggested indicators in this paper can capture different time scales, and thus allows for the trend detection.

In this research, Sina Microblog and 31 major cities in China were selected to explore the connectivity between cities in the spatial social network weaved by the human perceptions in the social media. A framework of measurements is proposed to estimate urban system connectedness in terms of asymmetric, relatedness, impact, dependence and structural similarity. The spatial distribution of these indicators can be further calculated by measurements such as Moran's *I*.

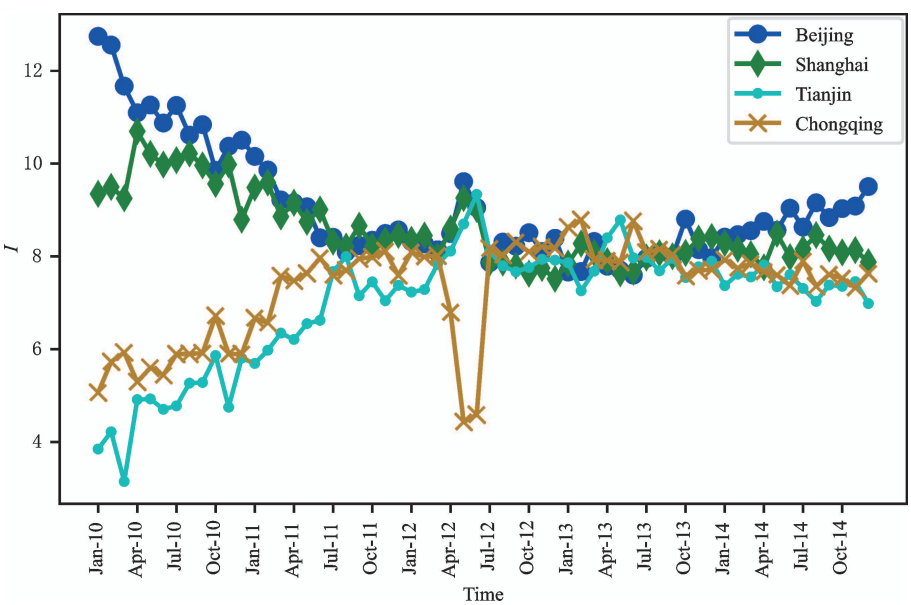

Fig. 9 Trend of impact index (*I*) for four cities in 2014

The study found that it is possible to measure the asymmetrical relatedness between geographical entities based on the number of references in social media. The strength of the link is significant correlated with the joint factor of population and GDP, and is influenced less by the distance when compared to the traditional gravity model. The asymmetrical relationship of impact/be impacted can be represented by the impact and dependence index, especially since the impact index is also the new comprehensive index of a city. Based on the asymmetrical relationship, the structural similarity index provides a new method, with high stability and consistency, to measure inter-city similarity. The indexes proposed in the paper can be applied to measure the relatedness between cities in different time scales and time intervals.

There are some limitations in using social media for connectivity analytics. First, the same location might not use the same toponyms. Due to the limitation of post length, there are numerous irregular abbreviations of place names, leading to a lack of standardization. This means that a part of the toponyms is not accurate and there are ambiguities (Lieberman and Samet, 2012). For example, Beijing might be called Royal Park, and California may be abbreviated to CA or Cal. Second, extreme events or breaking news would largely attract users' attention to specific cities, which might distort the city connectivity especially the posts are not collected in a long time period. This will affect the accuracy and consistency of measurement results. Third, the effects of ease of internet access and governmental policies upon social media will also affect the reliability of such measurements. In follow-up studies, other flow data will also be applied to validate our model. Semantic analysis can be used to discover both why and how asymmetric connectivity exists between cities.

References

Beaverstock J V, Smith R G, Taylor P J, 2000. World-city network: a new metageography? *Annals of the Association of American Geographers*, 90(1): 123–134. doi: 10.1111/0004-5608.00188

Borges K A V, Davis Jr C A, Laender A H F et al., 2011. Ontology-driven discovery of geospatial evidence in web pages. *Geoinformatica*, 15(4): 609–631. doi: 10.1007/s10707-010-0118-z

Buyukkokten O, Cho J, Garcia-Molina H et al., 1999. Exploiting geographical location information of web pages. In *ACM SIGMOD Workshop on the Web and Databases*. Philadelphia, Pennsylvania, ACM

Chen X, Yang X N, 2014. Does food environment influence food choices? A geographical analysis through 'tweets'. *Applied Geography*, 51: 82–89. doi: 10.1016/j.apgeog.2014.04.003

Erlander S, 1980. *Optimal Spatial Interaction and the Gravity Model*. Berlin: Springer. doi: 10.1007/978-3-642-45515-5

Gao S, Liu Y, Wang Y L et al., 2013. Discovering spatial interaction communities from mobile phone data. *Transactions in GIS*, 17(3): 463–481. doi: 10.1111/tgis.12042

Ghosh D D, Guha R, 2013. What are we 'tweeting' about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartography and Geographic Information Science*, 40(2): 90–102. doi: 10.1080/15230406.2013.776210

Ghufran M, Quercini G, Bennacer N, 2015. Toponym disambiguation in online social network profiles. *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*. Bellevue: ACM. doi: 10.1145/2820783.2820812

Grajales III F J, Sheps S, Ho K et al., 2014. Social media: a review and tutorial of applications in medicine and health care. *Journal of Medical Internet Research*, 16(2): e13. doi: 10.2196/jmir.2912

Han S Y, Tsou M H, Clarke K C, 2015. Do global cities enable global views? Using twitter to quantify the level of geographical awareness of U.S. cities. *PLoS One*, 10(7): e0132464. doi: 10.1371/journal.pone.0132464

Hay S I, George D B, Moyes C L et al., 2013. Big data opportunities for global infectious disease surveillance. *PLoS Medicine*, 10(4): e1001413. doi: 10.1371/journal.pmed.1001413

Haynes K E, Fotheringham A S, 1984. *Gravity and Spatial Interaction Models*. Beverly Hills: SAGE Publication.

Hecht B, Moxley E, 2009. Terabytes of tobler: evaluating the first law in a massive, domain-neutral representation of world knowledge. *Proceedings of the 9th International Conference on Spatial Information Theory*. Berlin: Springer. doi: 10.1007/978-3-642-03832-7_6

Hu Y J, Ye X Y, Shaw S L, 2017. Extracting and analyzing semantic relatedness between cities using news articles. *International Journal of Geographical Information Science*, 31(12): 2427–2451. doi: 10.1080/13658816.2017.1367797

Jiang B, Jia T, 2011. Exploring human mobility patterns based on location information of US flights. arXiv preprint, arXiv: 1104.4578. https://arxiv.org/abs/1104.4578

Jiang Y Q, Li Z L, Ye X Y, 2018. Measuring inter-city network using digital footprints from twitter users. *Proceedings of the 2nd ACM SIGSPATIAL Workshop on Prediction of Human Mobility*. Seattle: ACM. doi: 10.1145/3283590.3283594

Jones C B, Purves R S, 2008. Geographical information retrieval. *International Journal of Geographical Information Science,* 22(3): 219–228. doi: 10.1080/13658810701626343

Jones C B, Purves R S, Clough P D et al., 2008. Modelling vague places with knowledge from the Web. *International Journal of Geographical Information Science*, 22(10): 1045–1065. doi: 10.1080/13658810701850547

King D, Ramirez-Cano D, Greaves F et al., 2013. Twitter and the

health reforms in the English National Health Service. *Health Policy*, 110(2–3): 291–297. doi: 10.1016/j.healthpol.2013.02.005

Kitamoto A, Sagara T, 2012. Toponym-based geotagging for observing precipitation from social and scientific data streams. *Proceedings of ACM Multimedia 2012 Workshop on Geotagging and its Applications in Multimedia*. Nara, Japan: ACM. doi: 10.1145/2390790.2390799

Li S W, Ye X Y, Lee J et al., 2017. Spatiotemporal analysis of housing prices in China: a big data perspective. *Applied Spatial Analysis and Policy*, 10(3): 421–433. doi: 10.1007/s12061-016-9185-3

Lieberman M D, Samet H, 2012. Adaptive context features for toponym resolution in streaming news. *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Portland: ACM. doi: 10.1145/2348283.2348381

Lin J Y, Li X, 2015. Simulating urban growth in a metropolitan area based on weighted urban flows by using web search engine. *International Journal of Geographical Information Science*, 29(10): 1721–1736. doi: 10.1080/13658816.2015.1034721

Liu Y, Sui Z W, Kang C G et al., 2014b. Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. *PLoS One*, 9(1): e86026. doi: 10.1371/journal.pone.0086026

Liu Y, Wang F H, Kang C G et al., 2014a. Analyzing relatedness by toponym co-occurrences on web pages. *Transactions in GIS*, 18(1): 89–107. doi: 10.1111/tgis.12023

Luo W, Hartmann J F, Wang F H, 2010. Terrain characteristics and Tai toponyms: a GIS analysis of *Muang*, *Chiang* and *Viang*. *GeoJournal*, 75(1): 93–104. doi: 10.1007/s10708-009-9291-8

Massa P, Campagna M, 2014. Social media geographic information: recent findings and opportunities for smart spatial planning. *TeMA-Journal of Land Use*, *Mobility and Environment*. doi: 10.6092/1970-9870/2500

Miller H J, 2004. Tobler's first law and spatial analysis. *Annals of the Association of American Geographers*, 94(2): 284–289. doi: 10.1111/j.1467-8306.2004.09402005.x

Overell S, Rüger S, 2008. Using co-occurrence models for placename disambiguation. *International Journal of Geographical Information Science*, 22(3): 265–287. doi: 10.1080/13658810701626236

Padmanabhan A, Wang S W, Cao G F et al., 2014. FluMapper: a cyberGIS application for interactive analysis of massive location‐based social media. *Concurrency and Computation: Practice and Experience*, 26(13): 2253–2265. doi: 10.1002/cpe.3287

Porojan A, 2001. Trade flows and spatial effects: the gravity model revisited. *Open Economies Review*, 12(3): 265–280. doi: 10.1023/A:1011129422190

Sakaki T, Okazaki M, Matsuo Y, 2013. Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Transactions on Knowledge and Data Engineering*, 25(4): 919–931. doi: 10.1109/TKDE.2012.29

Samet H, Sankaranarayanan J, Lieberman M D et al., 2014. Reading news with maps by exploiting spatial synonyms. *Communications of the ACM*, 57(10): 64–77. doi: 10.1145/2629572

Sanderson M, Kohler J, 2004. Analyzing geographic queries. *Proceedings of the 27th Annual International ACM SIGIR Conference*. *Sheffield, UK : ACM*

Scellato S, Noulas A, Mascolo C, 2011. Exploiting place features in link prediction on location-based social networks. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Diego, USA: ACM. doi: 10.1145/2020408.2020575

Tobler W R, 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(S1): 234–240. doi: 10.2307/143141

Tsou M H, Yang J A, Lusher D et al., 2013. Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing): a case study in 2012 US Presidential Election. *Cartography and Geographic Information Science*, 40(4): 337–348. doi: 10.1080/15230406.2013.799738

Velardi P, Stilo G, Tozzi A E et al., 2014. Twitter mining for fine-grained syndromic surveillance. *Artificial Intelligence in Medicine*, 61(3): 153–163. doi: 10.1016/j.artmed.2014.01.002

Wang F H, Wang G X, Hartmann J et al., 2012. Sinification of Zhuang place names in Guangxi, China: a GIS-based spatial analysis approach. *Transactions of the Institute of British Geographers*, 37(2): 317–333. doi: 10.1111/j.1475-5661.2011.00471.x

Wang F H, Zhang L, Zhang G P et al., 2014. Mapping and spatial analysis of multiethnic toponyms in Yunnan, China. *Cartography and Geographic Information Science*, 41(1): 86–99. doi: 10.1080/15230406.2013.831529

Wang X G, Zhang Y, Chen M et al., 2010. An evidence-based approach for toponym disambiguation. *Proceedings of the 2010 18th International Conference on Geoinformatics*. Beijing: IEEE. doi: 10.1109/GEOINFORMATICS.2010.5567805

Wasserman S, Faust K, 1994. *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

Widener M J, Li W W, 2014. Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US. *Applied Geography*, 54: 189–197. doi: 10.1016/j.apgeog.2014.07.017

Wu J S, Feng Z, Zhang X W et al., 2020. Delineating urban hinterland boundaries in the Pearl River Delta: an approach integrating toponym co-occurrence with field strength model. *Cities*, 96: 102457. doi: 10.1016/j.cities.2019.102457

Ye X Y, Li S W, Yang X N et al., 2018. The fear of Ebola: a tale of two cities in China. In: Shen Z J, Li M Y (eds). *Big Data Support of Urban Planning and Management*. Cham: Springer, 113–132. doi: 10.1007/978-3-319-51929-6_7

Zhang W, Gelernter J, 2014. Geocoding location expressions in Twitter messages: a preference learning method. *Journal of Spatial Information Science*, 9(9): 37–70. doi: 10.5311/JOSIS.2014.9.170

Zhen F, Qin X, Ye X Y et al., 2019. Analyzing urban development patterns based on the flow analysis method. *Cities*, 86: 178–197. doi: 10.1016/j.cities.2018.09.015