Physica A 466 (2017) 462–475

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Analysis of co-occurrence toponyms in web pages based on complex networks

Xiang Zhong, Jiajun Liu, Yong Gao [*], Lun Wu

Institute of Remote Sensing and Geographic Information Systems, Peking University, Beijing 100871, China

HIGHLIGHTS

- Proposed the toponym co-occurrence network model.
- Found characteristics of this toponym co-occurrence network, including small world, scale-free and disassortative.
- Presented two new methods to extract core toponyms from web pages.

ARTICLE INFO

Article history:
Received 7 January 2016
Received in revised form 8 July 2016
Available online 3 October 2016

Keywords:
Toponym
Toponym co-occurrence
Complex network
Link analysis
Geographic information retrieval

ABSTRACT

A large number of geographical toponyms exist in web pages and other documents, providing abundant geographical resources for GIS. It is very common for toponyms to co-occur in the same documents. To investigate these relations associated with geographic entities, a novel complex network model for co-occurrence toponyms is proposed. Then, 12 toponym co-occurrence networks are constructed from the toponym sets extracted from the People's Daily Paper documents of 2010. It is found that two toponyms have a high co-occurrence probability if they are at the same administrative level or if they possess a part-whole relationship. By applying complex network analysis methods to toponym co-occurrence networks, we find the following characteristics. (1) The navigation vertices of the co-occurrence networks can be found by degree centrality analysis. (2) The networks express strong cluster characteristics, and it takes only several steps to reach one vertex from another one, implying that the networks are small-world graphs. (3) The degree distribution satisfies the power law with an exponent of 1.7, so the networks are free-scale. (4) The networks are disassortative and have similar assortative modes, with assortative exponents of approximately 0.18 and assortative indexes less than 0. (5) The frequency of toponym co-occurrence is weakly negatively correlated with geographic distance, but more strongly negatively correlated with administrative hierarchical distance. Considering the toponym frequencies and co-occurrence relationships, a novel method based on link analysis is presented to extract the core toponyms from web pages. This method is suitable and effective for geographical information retrieval.

1. Introduction

The Internet has become an important tool by which to deliver and exchange information, undergoing rapid development and attaining high popularity, and contains large amounts of geospatially referenced information. Location information is often associated with these web pages, such as news articles and blog posts [1]. Meanwhile, individuals can contribute

[*] Corresponding author.
E-mail address: gaoyong@pku.edu.cn (Y. Gao).

http://dx.doi.org/10.1016/j.physa.2016.09.024
0378-4371/© 2016 Elsevier B.V. All rights reserved.

geographical information to the web by sharing their positions, which brings enormous opportunities as well as challenges for the deep mining and knowledge discovery of geographical information [2–4].

Toponyms are among the most common types of geographical information in web documents [5]. The term refers to a proprietary name for a geographic entity with a specific orientation and a special geographical area. It can also be defined as a description of a location, which is a predetermined place [6]. Toponyms can generally be found in newspapers, travel notes, and so on. If two toponyms appear in the same web page, they are co-occurring. Co-occurring toponyms have strong relevance (such as political, economic and spatial relation) when they appear in the same topic of web texts. The more frequently one pair of toponyms co-occurs, the stronger the relevance between them. Co-occurring toponyms are first applied to dissolve the ambiguity of geographical names, namely, in toponym disambiguation [7–10], and are also used to research geographical relatedness. Liu et al. present a method to capture the relatedness between geographical entities based on their occurrences on web pages [11]. Analyzing the co-occurrence and topological distance between all pairs of geographical entities indicates that spatially close toponyms generally have similar co-occurrence patterns, and the frequency of co-occurrence exhibits a distance decay effect under the power law.

The co-occurrence phenomenon was first proposed in natural language processing to study word distribution. Word co-occurrence is defined as several words appearing together with a certain frequency in the same text document. Co-occurrence is widely applied in linguistics, scientific research cooperation and other fields, and has made laudable achievements. Brian Roark et al. present an algorithm for extracting potential entries for a category from an on-line corpus, based on a small set of exemplars [12]. Liang et al. construct a Chinese character (word) co-occurrence network based on the usage features of Chinese characters in poems and analyze the overall structure characteristics of poem co-occurrence networks from a linguistic perspective [13–15]. White and Griffith investigate the cooperation relationships of co-authors based on the number of papers the authors cite together, and they offer a new technique that contributes to the understanding of intellectual structure in science and possibly in other areas [16]. Leydesdorff et al. present both a global map with the functionality of a Google Map (e.g., zooming) and network maps based on normalized relations [17], and Qiu et al. provide a better comprehension of author interaction and contribute to the cognitive application of author co-occurrence network analysis [18]. Henry Small proposes a method to measure the relationship between two co-citation documents and provides a new approach to the study of SDI profiles [19]. Moreover, Rada Mihalcea extracts the key words and sentences from documents based on considering the frequency of co-words, which could automatically create an index for document collection [20].

In addition to many toponyms co-occurring in the same document, a toponym may also appear in many documents simultaneously. As a result, co-occurring toponyms are connected by a document, and documents are associated when they share the same toponym. The iterative associations will further lead to indirect connections between toponyms and finally form a network structure. Therefore, a novel complex network model is proposed in this paper to model the co-occurrence of toponyms. Twelve sample co-occurrence networks of Chinese geographical names are constructed and analyzed to investigate their structural characteristics, including centrality, degree distributions, the small-world feature and assortativeness. Furthermore, a link-based method is presented to find core toponyms, which can help to extract rich geographical information from massive web documents.

This paper is organized as follows. Section 2 proposes the co-occurrence toponym network model, using the frequency and the co-occurrence relationship of geographic names. Section 3 constructs 12 sample co-occurrence networks and analyzes their structural features. Section 4 gives an effective link-based method to obtain the core toponyms from the web page collection, and experiments are conducted to verify this technical solution. The discussions and conclusions are drawn in Sections 5 and 6.

2. Modeling toponym co-occurrence networks

Thousands of web pages form the major sources in the Internet. Web pages are organized by a document collection, which are represented as

$$W = \{D_1, D_2, \ldots, D_n\}. \tag{1}$$

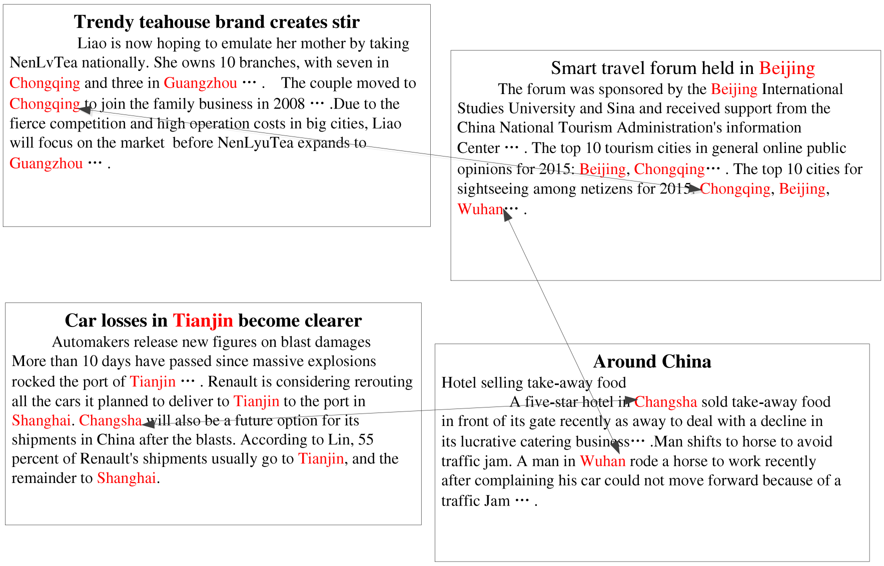
A web page is a collection of toponyms, represented as

$$D_i = \{x_1, x_2, \ldots, x_n\} \tag{2}$$

where a toponym is represented as x_i.

The appearance of two toponyms in the same document is defined as toponym co-occurrence. Thus, every two toponyms in a document, i.e., $\forall x_p, x_q \in D_i$, are co-occurring. A toponym participating in co-occurrence might also exist in many other documents, so these documents could be connected indirectly by sharing toponyms. A strong interrelated correlation between co-occurrence toponyms shows when they appear together in web pages. Considering the co-occurrence relationship and the transmission effect of toponyms, a graph structure is constructed, namely a toponym co-occurrence network. In this network, the toponyms extracted from web pages form the vertices of the graph, and their co-occurrence relationships are expressed by the edges. Two toponyms are linked by an edge if they co-occur in the same web page.

Formally, let $G = (V, E)$ be an undirected graph in which the out-degree of a vertex equals the in-degree of the vertex. The defined graph contains a set of vertices V and a set of edges E, where E is a subset of $V \times V$. The number of vertices N is



Trendy teahouse brand creates stir

Liao is now hoping to emulate her mother by taking NenLvTea nationally. She owns 10 branches, with seven in Chongqing and three in Guangzhou ⋯ . The couple moved to Chongqing to join the family business in 2008 ⋯ .Due to the fierce competition and high operation costs in big cities, Liao will focus on the market before NenLyuTea expands to Guangzhou ⋯ .

Smart travel forum held in Beijing

The forum was sponsored by the Beijing International Studies University and Sina and received support from the China National Tourism Administration's information Center ⋯ . The top 10 tourism cities in general online public opinions for 2015: Beijing, Chongqing⋯ . The top 10 cities for sightseeing among netizens for 2015: Chongqing, Beijing, Wuhan⋯ .

Car losses in Tianjin become clearer

Automakers release new figures on blast damages More than 10 days have passed since massive explosions rocked the port of Tianjin ⋯ . Renault is considering rerouting all the cars it planned to deliver to Tianjin to the port in Shanghai. Changsha will also be a future option for its shipments in China after the blasts. According to Lin, 55 percent of Renault's shipments usually go to Tianjin, and the remainder to Shanghai.

Around China

Hotel selling take-away food

A five-star hotel in Changsha sold take-away food in front of its gate recently as away to deal with a decline in its lucrative catering business⋯ .Man shifts to horse to avoid traffic jam. A man in Wuhan rode a horse to work recently after complaining his car could not move forward because of a traffic Jam ⋯ .

Fig. 1. Toponym relationships in web pages.

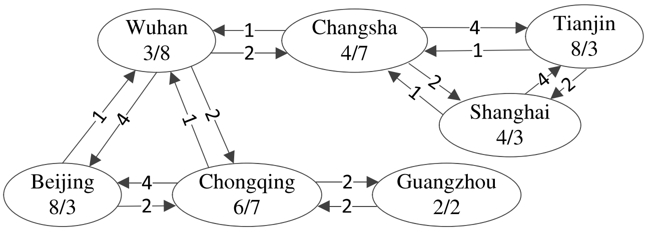


Fig. 2. Weighted directed graph for co-occurring toponyms.

equal to $|V|$, and the edge number is represented by $M = |E|$. G can be denoted by a co-occurrence matrix A with elements corresponding to the relative frequency of occurrence. In the corresponding matrix A, $\forall v_i, v_j \in V$, if $e_{ij} \in E$, then $a_{ij} = 1$; otherwise $a_{ij} = 0$. A is a symmetric matrix, i.e., $\forall v_i, v_j \in V, e_{ij} \in E, e_{ij} = e_{ji}$.

The weight of each edge is denoted by the frequency of co-occurrence. Supposing that x_1 and x_2 appear c_1 and c_2 times in a web page, respectively, the co-occurrence relation is manifested by an edge of the directed graph. Normally, x_1 refers to x_2 if x_1 points to x_2, and the weight of the corresponding edge is c_2. Similarly, the weight of the edge if x_2 points to x_1 is c_1. Thus, the co-occurrence network transforms to a weighted directed graph $G = (V, E, W)$ with a set of vertices V, a set of edges E, and a set of weights W. Each edge is represented as a triple $e_{ij} = \left(v_i, v_j, w_{ij}\right)$, where v_i is the starting point, v_j is the ending point, and w_{ij} is the weight, which equals the frequency of v_j. The calculation formula of w_{ij} is as follows:

$$w_{ij} = \sum_{v_i \in B(v_j)}^{n} c_j \tag{3}$$

where $B(v_j)$ is the set of toponym vertices that have a co-occurrence relation with v_j, and c_j is the frequency of v_j in the web page D_j.

For example, there is a set of web pages shown in Fig. 1. The set consists of 4 web pages, and each page contains several toponyms. Formally, the sets of toponyms are represented as $\{$(Changsha^1, Tianjin^4, Shanghai^2), (Wuhan^1, Changsha^1), (Beijing^4, Wuhan^1, Chongqing^2) and (Guangzhou^2, Chongqing^2)$\}$, in which the superscript number of each toponym is its frequency in the corresponding page. Fig. 2 shows the weighted directed graph constructed symmetric matrix, i.e., the toponyms from the page set. As an example, "Changsha" appears both in document 1 (Changsha^2, Tianjin^6, Shanghai^8) and document 2 (Wuhan^1, Changsha^2), so the two documents connect directly, and the remaining toponyms in the two documents are connected indirectly by "Changsha".

3. Structural analysis of toponym co-occurrence network

To investigate the structural properties of the toponym co-occurrence network, news websites are utilized as a corpus to construct sample networks. In this research, all of the news from People's Daily of China in 2010 is collected using

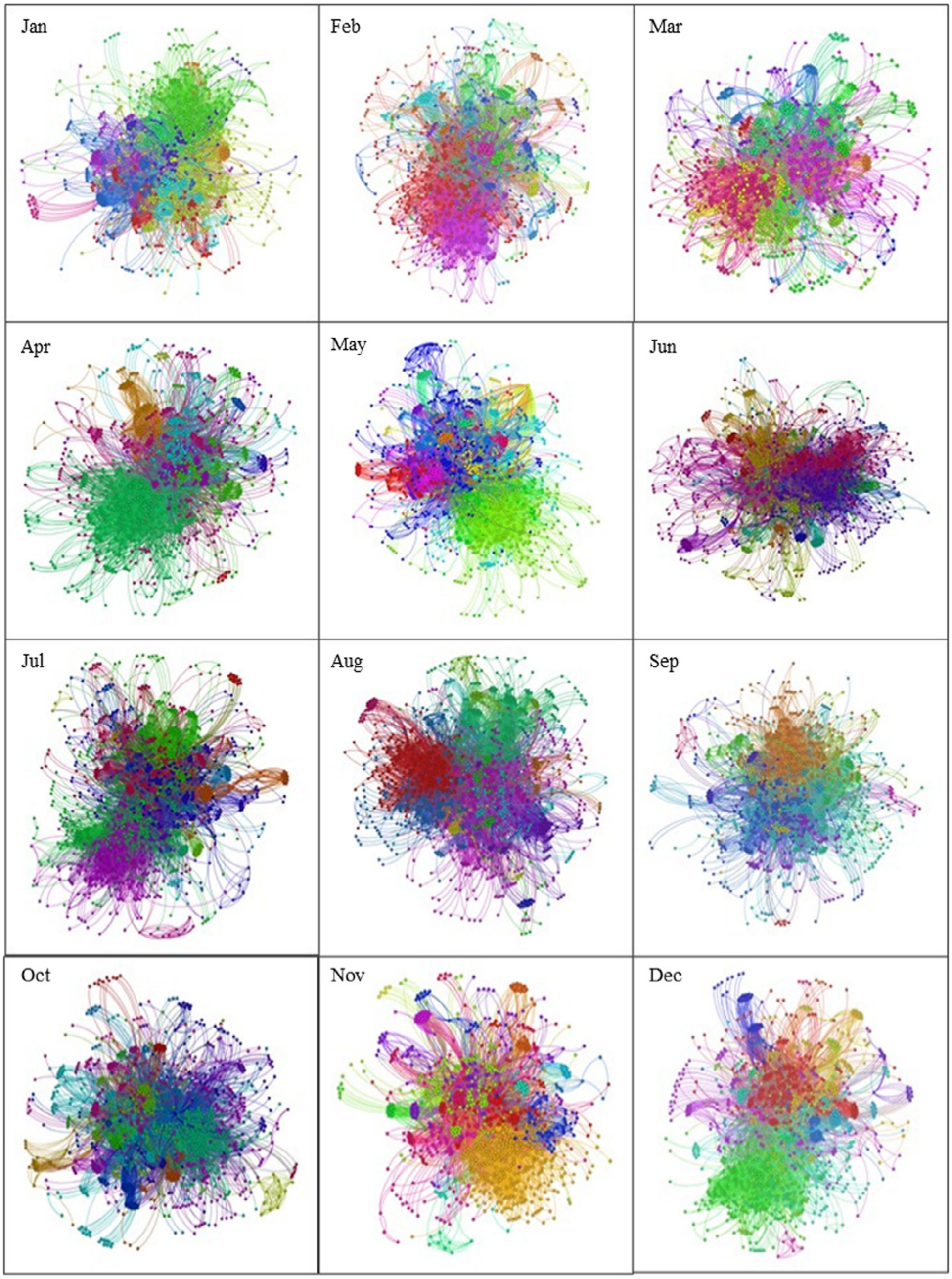


Fig. 3. Toponym co-occurrence networks constructed from People's Daily of China in 2010.

a vertical web crawler program [21], then organized in units of months. To ensure the data quality and improve the efficiency and accuracy of the experiment, we first preprocess the raw data by deleting documents that contain no toponyms. Then, the Language Technology Platform [22], an open-source Chinese natural language processing tool, is used to extract toponyms from the web page text. Finally, 12 toponym co-occurrence networks are constructed (Fig. 3) from the toponym sets.

Networkx [23], a python package for complex network analysis [24], is used to obtain the statistical structure parameters of these networks, and the results are shown in Table 1.

To investigate the impact of scale on structural features in the network model, more experiments are conducted for comparative analysis, using data from one day, one week, half a month, one month, three months, six months and one year. Table 2 illustrates that all of the statistical parameters stay stable, except that $\langle k\rangle$ increases with the network scale. For convenience, the following chapters were conducted with the data from People's Daily of China of twelve months in 2010.

Table 1
Statistical structure parameters of toponym co-occurrence networks.

Month	$\|V\|$	$\|E\|$	$\langle k\rangle$	γ	L	L_r	$C\%$	$C_r\%$	μ	Γ
1	1105	13 696	24.79	1.684	2.51	2.54	83.15	2.16	0.1355	−0.1012
2	1192	14 267	23.93	1.690	2.61	2.59	85.60	2.07	0.1730	−0.0818
3	808	7 467	18.48	1.700	2.54	2.62	81.99	2.24	0.1588	−0.1247
4	1108	14 780	26.67	1.675	2.50	2.49	84.98	2.33	0.1166	−0.1168
5	1186	12 155	20.50	1.700	2.50	2.68	82.90	1.79	0.2196	−0.1153
6	980	8 735	17.82	1.709	2.68	2.7	82.56	1.89	0.1564	−0.1037
7	1264	12 911	20.43	1.698	2.62	2.7	82.70	1.62	0.1964	−0.1067
8	1327	14 485	21.83	1.689	2.62	2.67	83.40	1.63	0.1646	−0.0880
9	1031	9 628	18.67	1.711	2.50	2.69	81.40	1.82	0.2547	−0.1256
10	1140	12 149	21.31	1.693	2.51	2.64	83.80	1.89	0.2091	−0.1082
11	966	10 269	21.26	1.691	2.42	2.59	83.17	2.22	0.2400	−0.1168
12	1131	12 274	21.70	1.685	2.55	2.63	84.30	1.91	0.1848	−0.0772

Table 2
Statistical structure parameters of toponym co-occurrence networks.

Scale	$\|V\|$	$\|E\|$	$\langle k\rangle$	γ	L	L_r	$C\%$	$C_r\%$	μ	Γ
One day	75	401	10.69	1.6445	2.18	2.04	0.91	0.15	0.096	−0.19
One week	338	2 303	13.63	1.7206	2.48	2.52	0.86	0.04	0.165	−0.13
Half a month	571	4 489	15.72	1.7145	2.5	2.61	0.84	0.028	0.252	−0.13
One month	1105	13 696	24.79	1.6643	2.51	2.54	0.83	0.022	0.136	−0.11
Three months	2282	30 918	27.10	1.6775	2.57	2.71	0.84	0.012	0.168	−0.12
Six months	3869	56 532	29.22	1.6839	2.57	2.79	0.84	0.008	0.219	−0.13
One year	6627	103 210	31.15	1.6906	2.57	2.87	0.84	0.0047	0.255	−0.13

3.1. Centrality analysis

In complex network theories, indicators of centrality identify the significance of vertices within a graph. Centrality analysis is widely applied to social network analysis (SNA) [25], especially to measure the importance of individuals. The importance of a vertex is reflected in its impact on other vertices or on the overall graph. Several centrality indications are proposed with different emphases, including Degree Centrality, Betweenness Centrality, Closeness Centrality, and so on [26].

(1) **Degree centrality**

The degree centrality of a toponym vertex is defined as the number of links incident upon this vertex. The more neighbors a toponym vertex has, the higher its degree. High degree indicates that the toponym has strong communication ability within the network because many other toponyms co-occur with it. The degree centrality of a toponym vertex is

$$C_{AD}(i) = \sum_{j=0,a_{ij}\neq 0}^{n} 1 \tag{4}$$

where a_{ij} represents the element at row i and column j of the network matrix. The result of normalized degree centrality analysis for the top 20 toponyms of January is shown in Table 2. Other months exhibit similar results. "China" has a maximal degree in this graph, as more than half of the other toponyms are its neighbors. The specific experimental data may be one of the reasons because most of the news in People's Daily of China reports about China. At the same time, Beijing, as the capital of China, is often associated with China in news reports, so "Beijing" also appears frequently. The following characteristics can be observed from the results (Table 3): (1) The larger the geographical area one toponym represents, the higher the degree of attention it has. For example, compared to "Qingdao" (a coastal city in eastern China), "Shandong" is a province located in the east of China. It not only has a larger area but also contains many cities, including "Qingdao". Thus, "Shandong" receives more attention and has a higher degree in the graph. (2) The economic development level has a promoting influence. The developed cities attract more attention than the developing cities. "Shanghai" (the most developed city in China) appears in more web pages than "Wuhan" (a city undergoing rapid development).

(2) **Betweenness centrality**

Betweenness is another centrality measure of a vertex within a graph. Betweenness centrality quantifies the times that a vertex acts as a bridge along the shortest path between two other vertices. Many vertices are not directly related to each other but are connected through other intermediate vertices. Betweenness is then applied to measure the connectivity potential of a toponym vertex. It reflects the ability of a vertex to control communication among vertices. The betweenness of a vertex is computed as follows:

$$B_{AD}(i) = \sum_{s,t\in V; s,t\neq i} \frac{\sigma(s,t|i)}{\sigma(s,t)} \tag{5}$$

Table 3
Degree centrality of the top 20 toponyms in January.

Rank	Toponym	Degree centrality	Rank	Toponym	Degree centrality
1	China	0.562	11	Hebei	0.142
2	Beijing	0.327	12	Anhui	0.141
3	America	0.274	13	France	0.135
4	Shanghai	0.219	14	Japan	0.133
5	Xinjiang	0.172	15	Brazil	0.126
6	Asia	0.165	16	Chongqing	0.126
7	Shandong	0.157	17	Yunnan	0.122
8	Guangdong	0.153	18	England	0.121
9	Russia	0.150	19	Liaoning	0.121
10	Europe	0.143	20	Henan	0.120

Table 4
Betweenness centrality of the top 20 toponyms in January.

Rank	Toponym	Betweenness centrality	Rank	Toponym	Betweenness centrality
1	China	0.35179	11	Sichuan	0.01947
2	Beijing	0.10958	12	Shandong	0.01907
3	America	0.08146	13	Inner Mongolia	0.01905
4	Shanghai	0.04515	14	Henan	0.01864
5	Xinjiang	0.04079	15	Hebei	0.01859
6	Anhui	0.02715	16	Asia	0.01656
7	Chongqing	0.02427	17	Guangzhou	0.01407
8	Shanxi	0.02382	18	Europe	0.01400
9	Guangdong	0.02204	19	Korea	0.01393
10	Russia	0.02126	20	Chengdu	0.01385

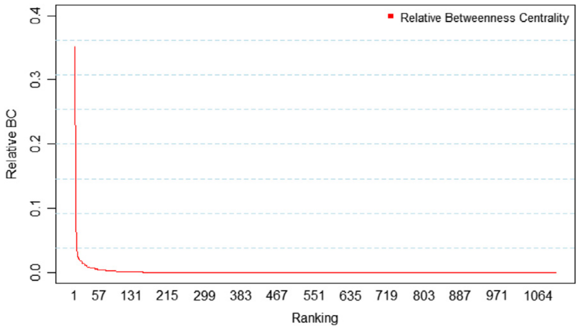


Fig. 4. The rank distribution of betweenness centrality.

where s and t are a pair of vertices other than vertex i, $\sigma(s, t|i)$ is the number of times that vertex i acts as a bridge along the shortest path between s and t, and $\sigma(s, t)$ is the number of shortest paths between s and t.

From the results of Table 4, we find that a vertex with larger degree centrality usually has a larger betweenness centrality. This result demonstrates that the toponym co-occurrence network has a loose structure, and most vertices are connected to some focused vertices. These focused toponyms always appear in many documents and co-occur with some other toponyms. In comparison, marginal toponyms appear in fewer documents. With the rapid increase in the number of web pages, the scale of the co-occurrence network expands. The number of marginal vertices increases rapidly, and these marginal vertices can hardly reach each other without going through the focused vertices. The betweenness centrality measures the connectivity potential of a vertex. If these intermediate vertices were removed from the network, the connectivity of the entire network would suffer a terrible impact, some connections between vertices would be cut off, and the network would be separated into several irrelevant sub-networks. We find that "China" has the largest betweenness centrality, while the betweenness centrality of more than 60% of vertices in the network is 0. As shown in Fig. 4, the relative betweenness centrality distribution exhibits a heavy tailed distribution. Toponym vertices with large betweenness centralities are the most significant navigation vertices in the co-occurrence network. Ranking each toponym vertex with its betweenness centrality, we can extract the hub toponym set.

(3) **Closeness centrality**

In contrast to degree centrality, which emphasizes local connection characteristics, closeness centrality denotes the relative accessibility among vertices by considering the relative connectivity of vertices. Thus, its rank result measures the global centrality of a vertex, and it evaluates whether the vertex has an advantage in terms of spatial location based on the

Table 5
Closeness centrality of the top 20 toponyms in January.

Rank	Toponym	Closeness centrality	Rank	Toponym	Closeness centrality
1	China	0.69346	11	Russia	0.52050
2	Beijing	0.59195	12	Henan	0.52001
3	America	0.56384	13	Japan	0.51879
4	Shanghai	0.551448	14	Tibet	0.51855
5	Xinjiang	0.538011	15	Yunnan	0.51830
6	Guangdong	0.531535	16	Zhejiang	0.51612
7	Shandong	0.531280	17	Fujian	0.51396
8	Hebei	0.52571	18	Wenchuan	0.51372
9	Anhui	0.52546	19	Brazil	0.51277
10	Chongqing	0.52396	20	Sichuan	0.51182

distances to other vertices. It can be calculated by

$$R_{AD}(i) = \frac{1}{\sum\limits_{j=1,j\neq i}^{n} D_{ij}}. \tag{6}$$

Table 5 shows that "China" still has the largest closeness centrality, at 0.69, while the lowest is "Dongxing Town" with 0.26. The relative closeness centrality of more than half the vertices exceeds 0.5, which suggests a right-skewed distribution. Generally, a vertex has a larger closeness centrality when it is close to the "centroid" of the network. Moreover, compared to the marginal vertices, the vertices in the dense regions have higher closeness centralities. In summary, the closeness centrality distribution and degree centrality distribution have similar characteristics. A toponym vertex with a larger centrality always implies that its corresponding region has a wide geographical area or a high level of economic development.

3.2. Degree distribution

The degree distribution is the probability distribution of those degrees over the whole network and provides an effective analysis method of complex network analysis. The topology properties and kinetic behaviors of a complex network depend on the analysis of the degree distribution [27]. The degree distribution $p(k)$ of a network is defined as the fraction of vertices with degree k in the network. Thus, if there are N vertices in a network, and n_k of them have degree k, we have

$$p(k) = n_k/N. \tag{7}$$

However, our networks have a discrete degree distribution, which means that many degrees have a zero probability. The complementary distribution [28,29] is used to describe the degree distribution in our networks.

For the 12 toponym co-occurrence networks constructed from People's Daily of China (Fig. 3), Fig. 5 shows the relationship between the corresponding degree k and its complementary distribution probability $p(k)$. In each diagram, the horizontal axis shows the degree of the vertices, and the vertical axis is the corresponding complementary distribution probability. The blue dots denote the relationship of k and $p(k)$, and the red line is the fitting curve. The degrees of most vertices in these networks are less than 200. The larger the degree, the lower is the corresponding probability. All fitting curves satisfy the power law $p(k) \propto k^{-\gamma}$, where γ is a positive constant.

As shown in Tables 1 and 2, the value of γ falls into the range of 1.645–1.721, and the number of $\langle k \rangle$ remains constant in networks of the same scale but increases when the scale expands, which proves that these co-occurrence networks are scale-free [30,31]. Some focus vertices have a large number of neighbor vertices, while more vertices are on the verge of the graph.

3.3. Small-world feature

In a graph, the shortest path between any pair of vertices v_i and v_j is defined as the path passing through the least edges. Dijkstra [32] is the most famous and effective algorithm for solving the shortest-path problem.

Let d_{ij} be the length of the shortest path that connects two given vertices v_i and v_j. The average shortest path length of the network is defined as

$$L = 2\sum_{i>j} d_{ij}/N(N-1) \tag{8}$$

where N is the number of vertices, and L presents the degree of separation between vertices in the network. It is a global indicator of a network.

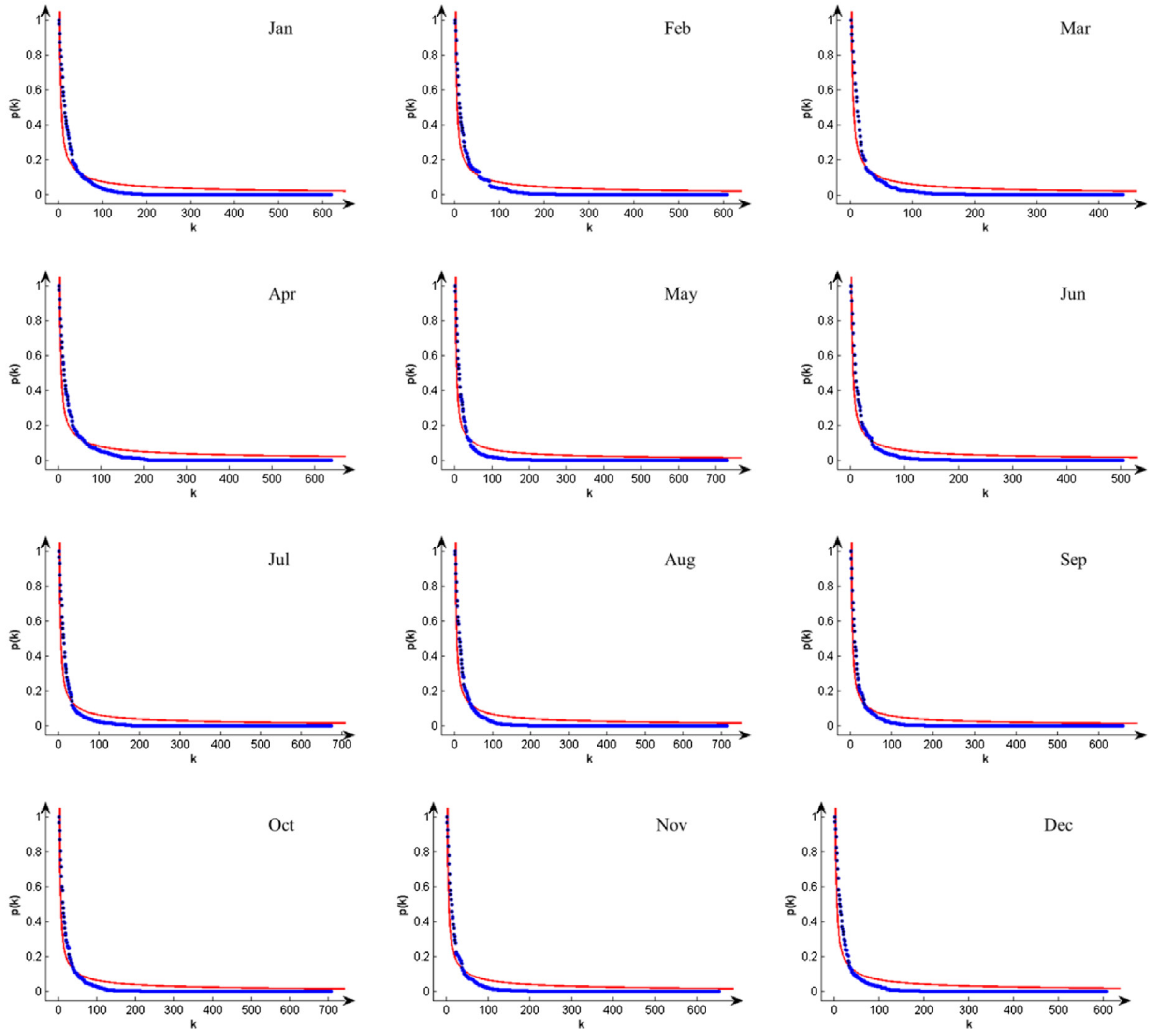


Fig. 5. The relationships between the degree k and the complementary distribution probability $p(k)$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The clustering coefficient C_i of the vertex v_i is the probability that any two neighbors of v_i are also connected to each other. Specifically, $C_i = 2E_i/k_i\,(k_i - 1)$, where E_i is the number of existing edges among the neighbors of v_i. The clustering coefficient of the whole network, namely C, is the average of C_i and reflects the local clustering feature of the network.

A network is a small-world network [33] if it has a small average shortest path length, while the clustering coefficient is higher than in random chance. In general, we can compare the toponym co-occurrence network G to a random graph GR [34] with the same number of vertices and edges. The toponym co-occurrence network exhibits small-world features when it satisfies the following two conditions: (1) $L \approx L_r$ and (2) $C \gg C_r$.

The results (the sixth and eighth columns in Table 1) tell us that the average shortest path length L of these toponym co-occurrence networks ranges from 2.42 to 2.68, with an average of 2.55, while the average shortest path length L_r of the corresponding random networks ranges from 2.49 to 2.7, with an average of 2.63. Fig. 6(a) shows the comparison between the shortest path length of the co-occurrence networks and the corresponding random networks. Fig. 6(b) depicts the comparison of the clustering coefficients. All of the clustering coefficients are over 0.8, which reflects a high aggregation effect, while the clustering coefficients of random networks range from 0.0162 to 0.0233. For the 12 co-occurrence networks, all of them have $L \approx L_r$ and $C \gg C_r$, exhibiting the small-world feature. The reason for this effect lies in the inherent structural features of toponym co-occurrence networks. In the networks, every edge represents the co-occurrence relationship between two toponyms, and the distance between co-occurring toponym vertices is 1. In Table 1, we find that the average degree of these co-occurrence networks is 24.5, meaning that every vertex has 24.5 neighbors in the networks. Thus, despite the large number of vertices in the networks, the average path length is very small.

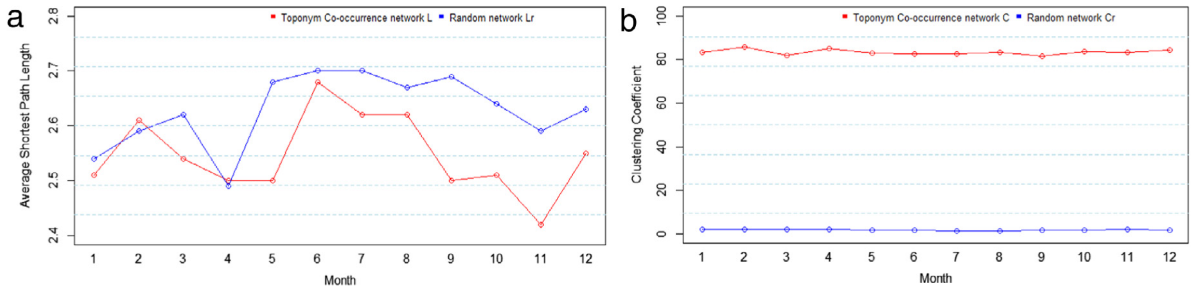


Fig. 6. Toponym co-occurrence networks and random networks with (a) average shortest path length comparison and (b) clustering coefficient comparison.

3.4. *Assortativeness analysis*

The formation mechanism of a network is described by its assortativeness. It is a preference for a network's vertices to connect to others that are similar in some way. The most prominent measures for capturing this feature are the assortativeness coefficient Γ [35] and the neighbor connectivity k_{nn} [36].

The assortativeness coefficient is the Pearson correlation coefficient of the degrees between pairs of linked vertices, and it is given by

$$\Gamma = \frac{E^{-1}\sum_i j_i k_i - \left[E^{-1}\sum_i \frac{1}{2}\left(j_i + k_i\right)\right]^2}{E^{-1}\sum_i \frac{1}{2}\left({j_i}^2 + {k_i}^2\right) - \left[E^{-1}\sum_i \frac{1}{2}\left(j_i + k_i\right)\right]^2} \tag{9}$$

where j_i and k_i are the degrees of the start and end vertices of edge i, and E is the number of edges in the network. In general, Γ lies between -1 and 1. When $\Gamma > 0$, the network has assortative mixing patterns, and vertices tend to be connected to other vertices with similar degree values; when $\Gamma = 0$, the network is non-assortative; and at $\Gamma < 0$, the network is disassortative, and the high-degree vertices tend to connect to low-degree vertices.

Another method of capturing the degree correlation is examining the properties of k_{nn}, or the average degree of neighbors of a vertex with degree k. This term is formally defined as

$$k_{nn}\left(k\right) = \sum_{k'} k' p(k'|k) \tag{10}$$

where $p(k'|k)$ is the conditional probability that an edge starting at a vertex with degree k ends at a vertex with degree k'. If $k_{nn}\left(k\right) \propto k^{-\mu}$, then μ is the assortative exponent.

We adopt the above two approaches to study the assortativeness of toponym co-occurrence networks, as shown in Fig. 7.

In each diagram in Fig. 7, the vertex degree k is depicted on the x-axis, while the y-axis represents k_{nn}, the average degree of neighbors of a corresponding vertex with degree k. The blue dots are the scatter distribution diagram of k and k_{nn}, and the red curve is the fitting curve. All of the networks satisfy the power law distribution $k_{nn}\left(k\right) \propto k^{-\mu}$, where $\mu > 0$, which means they are disassortative. At the same time, the disassortativeness exponent μ is within a small scope, ranging from 0.136 to 0.255. These networks have minor differences in disassortativeness and exhibit a similar disassortativeness pattern. The network disassortativeness shows that the high-degree vertices are always the hubs in the network, such as “Asia”, “Beijing” and so on. These hub vertices do not gather together but are discretely distributed in the network as the transit vertex set. In contrast, the low-degree vertices are always at the edges of the network.

We find in Table 1 that all disassortative coefficients are less than 0, ranging from -0.13 to -0.08. This result verifies that toponym co-occurrence networks have a disassortative feature from another perspective. Toponyms with lower degree tend to connect to the focused toponyms.

4. Extracting core toponyms based on the co-occurrence network

The importance of a toponym in a document can be expressed by its co-occurrence. On the one hand, toponym co-occurrence reflects a strong relationship between the pair of toponyms, which is always used to explain their spatial correlation. On the other hand, documents can link to each other through shared toponyms.

A large number of co-occurring toponyms occur in web pages, which are valuable for GIS research. Much effort has been made to extract the core toponyms from web documents. The best-known method is based on the ATF-PDF model [37], a statistical method referring to the frequency of geographical names. However, the ATF-PDF model ignores the relationships of toponyms, and thus an unsatisfactory rank result may occur when a document contains many duplicate toponyms.

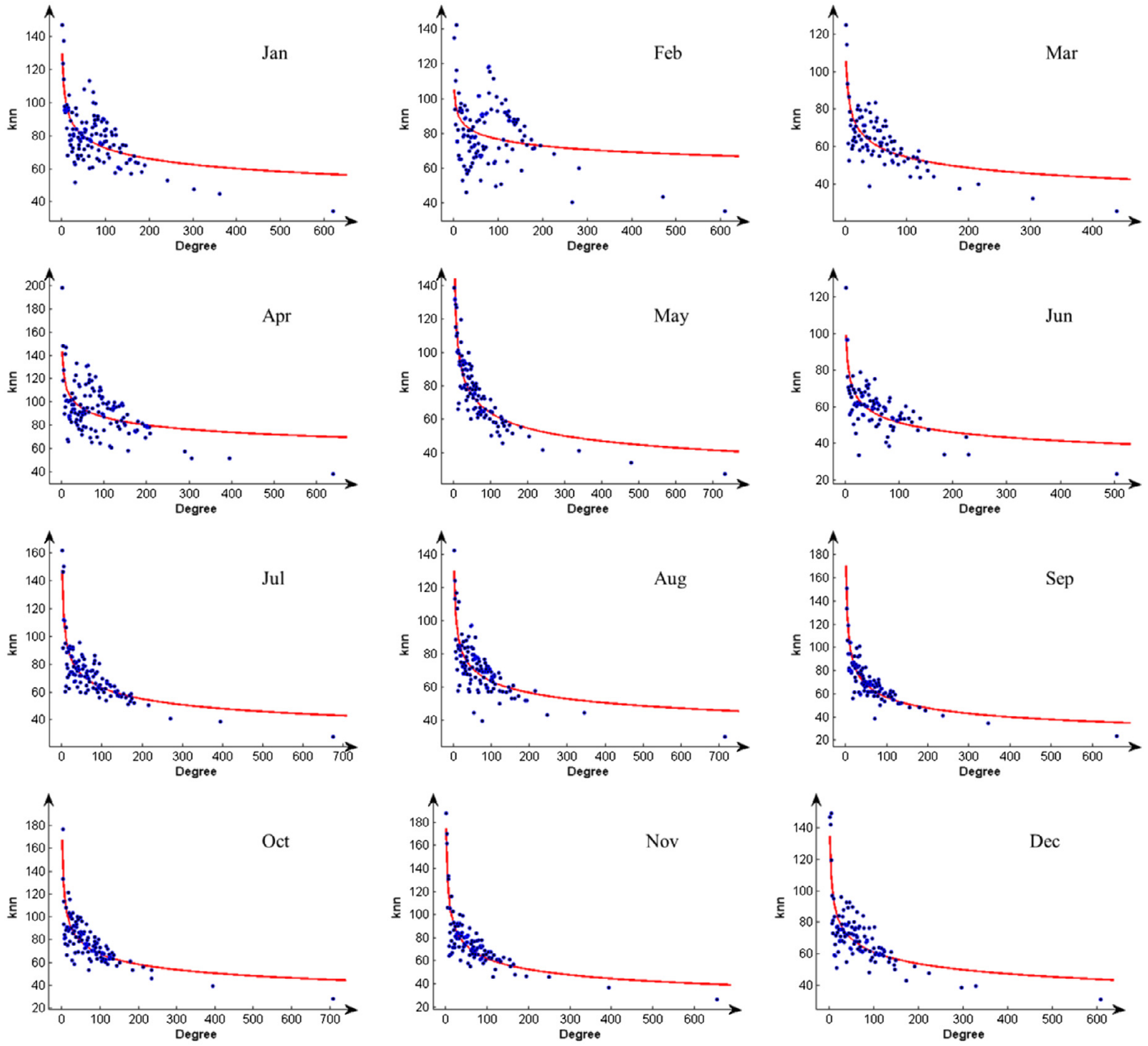


Fig. 7. Disassortativeness of the toponym co-occurrence networks. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In the toponym co-occurrence network, toponyms in the same document are directly connected by their co-occurrence relation, and different documents are connected by recurring toponyms. Thus, all toponyms have direct or indirect connections through the edges of the network. Toponym vertices with remarkable features or navigation functions in huge network resources are more important than the bulk of vertices. They are the core hubs in the entire network and have greater contributions to the connectivity and aggregation in the co-occurrence network. This factor can not only reflect the cognitive depth according to the frequency information but also represent the cognitive span with link relations.

Based on the co-occurrence network structure, we attempt to identify the core toponyms by using the toponym centrality, which considers the degree distribution of a network. A toponym vertex is more important than other vertices when it has more links. However, this method only considers the direct relations of toponyms, whereas the indirect relations and frequency features are lost.

Therefore, we propose a new link-based method that considers the toponyms' transition effect to extract the core toponyms by PageRank [38,39] based on toponym co-occurrence networks. PageRank, one of the best-known link analysis [40] algorithms for computing the principal eigenvector of the matrix of hyperlinks in web pages, is applied first in Google searches. The basic principle of the algorithm is to provide importance estimates for web pages via voting from other pages. A hyperlink is more popular when more links point to it, resulting in a higher rank. The convergence PageRank (*PR*) values of the vertices are obtained by the iterative calculation.

A transition rate matrix is established based on the weight of the vertices, and vertex scores are computed by the iterative PageRank calculation and called the PR values. The PR value of one vertex depends on the *PR* values of the other vertices it is related to. The *PR* value is computed as

$$PR\left(t_i\right) = \sum_{j \in B(t_i)}^{n} PR\left(t_j\right) \tag{11}$$

where $PR\left(t_i\right)$ denotes the ranking score of vertex t_i, and $B(t_i)$ are the set of other vertices that point to t_i.

To solve the problem that the *PR* value is zero when the hyperlink points to itself, a small damping factor d is added to every vertex. Hence, the final ranking formula of toponym vertices is updated using

$$PR\left(t_i\right) = d \sum_{j \in B(t_i)}^{n} \frac{PR(t_j)}{L(t_j)} + (1 - d)/N. \tag{12}$$

Generally, d is set as 0.85. The *PR* value of each vertex tends to be stabilized by iterative calculation even though they are set as a random non-zero value at first, and a convergence state can be reached. Then, the core toponyms of web pages will emerge according to the descending *PR* value arrangement.

The typical PageRank algorithm only focuses on the relationships of vertices, and each vote from other hyperlinks shares the same weight, ignoring the weights of edges, which have an important impact on the ranking result. Therefore, to satisfy the weighted vertex core ranking, the PageRank algorithm can be revised so that $L(t_j)$ represents the sum of the weights of the edges related to the vertex, not just the number of vertices to which the vertex points. Thus, the weighted core toponym *PR* is ultimately measured using

$$PR\left(t_i\right) = d \sum_{j \in B(t_i)}^{n} \frac{w_{ij} * PR(t_j)}{\mathrm{Sum}(t_j)} + (1 - d)/N \tag{13}$$

$$\mathrm{Sum}\left(t_j\right) = \sum_{i=0, i \neq j}^{n} w_{ij} \tag{14}$$

where $B(t_i)$ is the set of co-occurring toponyms, and N denotes the number of vertices. The higher the *PR* value of a vertex is, the more hub vertices point to it, which means that there are more toponyms co-occurring with it. When voted on by more toponyms, it achieves a higher rank in the toponym network.

Thus, by studying the transition effect between co-occurring toponyms, we iteratively calculate the importance value of a vertex based on link analysis, then identify the significant toponyms with remarkable features or navigation functions in huge network resources, which can be applied to geographical information retrieval.

Toponyms are extracted from People's Daily of China in January of 2010 and Sina Sports News in January of 2015, respectively. They are then ranked by the link-based method proposed and by ATF-PDF and centrality analysis for comparison. The results of the three methods are listed in Tables 6 and 7. The sorted results of the three methods are clearly very similar, which is because the high-frequency toponyms always co-occur with other toponyms. However, there is also a clear difference for some toponyms. The link-based method combines the advantages of the ADF-PDF method and centrality analysis method. Not only the indirect features of co-occurrence toponyms but also the frequency distributions are considered. In the toponym co-occurrence network constructed from Sina Sports News, there are many toponyms with a co-occurrence relationship with the word "Asia". This finding highlights the topic "Asia Cup", which was held in Australia at that time, attracting public attention. Taking advantage of this information, we can find the hot event based on the corresponding locations and apply the result to geographical information retrieval.

Obviously, the core toponyms extracted from web pages are specific to the theme of the corpus. A large number of news reports about the "Asia Cup" and "NBA" appeared in Sina Sports News of January, 2015, so we find that the competition sites, such as "Asia", "America" and "Australia", frequently appear in the sports news. These toponyms are in the core positions in the network and act as transmitters.

To conveniently discuss the changes in core toponyms in hot events with time, we use the news set of People's Daily of China from January to March and extract the top 10 ranked toponyms by the link-based method, as shown in Table 8. All of the news is under the same topic, but at different times. In January, the internal toponyms such as "Shanghai", "Xinjiang" and "Shandong" had higher rankings, while "Russia", "Japan", "Chile" and other international toponyms were in a leading position in March. This result reflects that the core toponyms change over time in co-occurrence networks, even within the same topic corpus. Because all co-occurrence relations are extracted from web documents, which are generally variable time, the core toponym results also vary with time. It could be an advantage to obtain real-time core toponyms based on co-occurrence methods. Thus, by extracting core toponyms from different time sequences, we could explore the variation of hot locations in hot events.

5. Discussion

A toponym is a description of a geographic feature or place in the real world. The relationship between co-occurring toponyms may be affected by the corresponding spatial relationships. Invoking the thumb law of geography, "Everything is

Table 6
Toponym ranking results for People's Daily News using ATF-PDF, degree centrality and the link-based method (top 10).

Ranking	ATF-PDF method		Degree centrality method		Link-based method	
	Toponym	ATF ∗ PDF	Toponym	Centrality (DC)	Toponym	PR
1	China	0.490583	China	0.562	China	0.0232
2	Beijing	0.10099	Beijing	0.327	Beijing	0.0131
3	America	0.055318	America	0.274	America	0.0099
4	Shanghai	0.039507	Shanghai	0.219	Shanghai	0.0086
5	Haiti	0.022525	Xinjiang	0.172	Xinjiang	0.0061
6	Xiaogang village	0.020617	Asia	0.165	Shandong	0.0055
7	Xinjiang	0.019691	Shandong	0.157	Guangdong	0.0054
8	Great Hall of the People	0.016037	Guangdong	0.153	Asia	0.0053
9	Iran	0.015747	Russia	0.150	Hebei	0.0052
10	Sichuan	0.013802	Europe	0.143	Russia	0.0051

Table 7
Toponym ranking results of Sina Sports news using ATF-PDF, degree centrality and the link-based method (top 10).

Ranking	ATF-PDF method		Degree centrality method		Link-based method	
	Toponym	ATF ∗ PDF	Toponym	Centrality (DC)	Toponym	PR
1	Beijing	0.3099	China	0.604	China	0.1078
2	China	0.2285	Beijing	0.589	Australia	0.0525
3	Australia	0.1139	Australia	0.330	Beijing	0.0417
4	Saudi Arabia	0.0457	Asia	0.305	Asia	0.0284
5	America	0.0450	Japan	0.295	Korea	0.0269
6	Uzbekistan	0.0404	Uzbekistan	0.293	Saudi Arabia	0.0263
7	Asia	0.0383	America	0.288	Uzbekistan	0.0263
8	Korea	0.0374	Korea	0.272	Japan	0.0262
9	Los Angeles	0.0363	Brazil	0.271	Brazil	0.0131
10	Brides-Bains	0.0357	Saudi Arabia	0.268	America	0.0124

Table 8
Link-based ranking results of samples in People's Daily from January to March (top 10).

Ranking	January		February		March	
	Toponym	PR	Toponym	PR	Toponym	PR
1	China	0.0232	China	0.0462	China	0.1461
2	Beijing	0.0131	Beijing	0.0240	Beijing	0.0591
3	America	0.0099	America	0.0217	Russia	0.0323
4	Shanghai	0.0086	Shanghai	0.0164	America	0.0276
5	Xinjiang	0.0061	Sichuan	0.0140	Tibet	0.0237
6	Shandong	0.0055	Tibet	0.0130	Shanghai	0.0171
7	Guangdong	0.0054	Russia	0.0122	Japan	0.0162
8	Asia	0.0053	Guangdong	0.0101	Chile	0.0124
9	Hebei	0.0052	Tianjin	0.0097	Xinjiang	0.0120
10	Russia	0.0051	Japan	0.0082	Taiwan	0.0112

related to everything else, but near things are more related than distant things" [41]. We have explored the relationships between the co-occurrence frequency of a pair of co-occurring toponyms and the corresponding geographic distance (Fig. 8(a)), as well as the average frequency along the distance (Fig. 8(b)). When the geographical distance is small (less than 500 km in these data, as shown in Fig. 8), the co-occurrence frequency shows a declining trend, representing a weak negative correlation. However, the correlation becomes undetectable when the geographical distance is long enough.

On the other hand, a toponym must correspond to a specific administrative unit. The relationship and interaction between two administrative districts in economics, culture or certain events can also produce the co-occurrence of their place names. Thus, we construct a tree structure according to the administrative hierarchy of China, which is organized in 5 levels, i.e., country, province, city, county and town. Any toponym has a node in this tree corresponding to the administrative unit it belongs to. The hierarchical distance of two toponyms is calculated by the shortest path between the two corresponding nodes in the tree. We investigate the relationship between the co-occurrence frequency and the corresponding hierarchical distance (Fig. 9) and find that the frequency is negatively correlated with the hierarchical distance. Two toponyms have a greater co-occurrence probability if they share a part–whole relationship, which is denoted as father and son in the tree; for example, Haidian is a district of Beijing. If two toponyms are in the same administrative division, depicted as a pair of siblings in the tree, e.g., Beijing and Shanghai, the probability is also higher.

In general, the co-occurrence of toponyms is affected more by administrative hierarchical relationships than spatial distance. The cultural and economic exchanges between different regions can overcome geographical separation, especially in the modern world. This phenomenon is more significant in news and other web documents which contain more political,

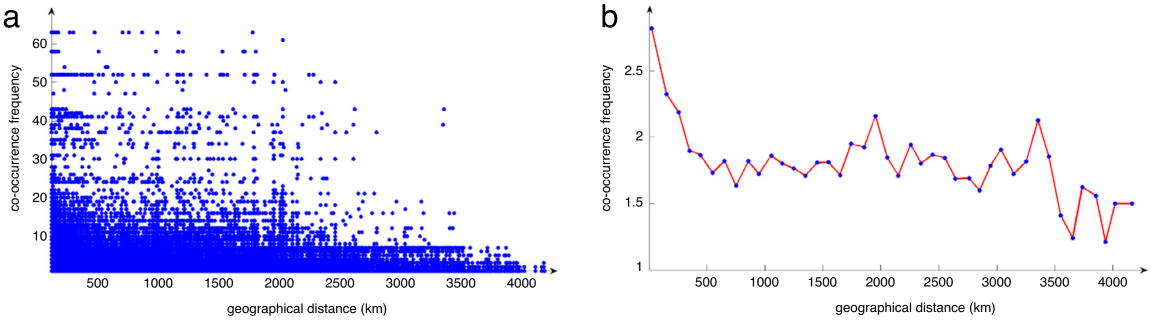


Fig. 8. Relationships between geographic distance and (a) co-occurrence frequency, (b) average co-occurrence frequency.

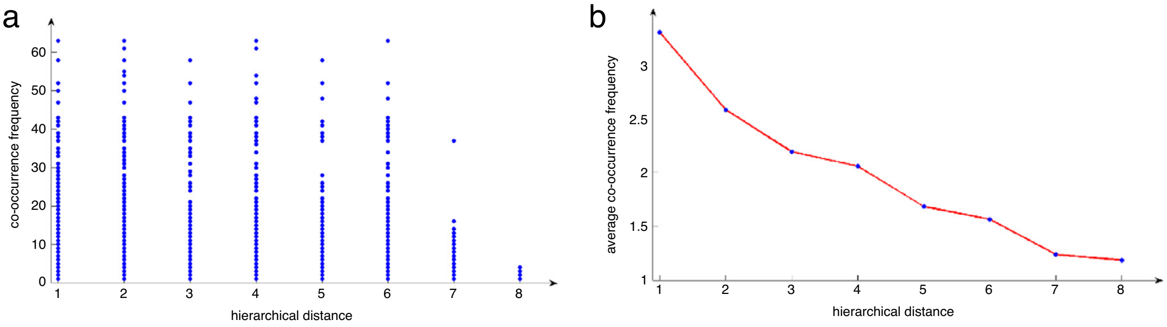


Fig. 9. Relationships between hierarchical distance and (a) co-occurrence frequency, (b) average co-occurrence frequency.

economic and life contents, which also demonstrates the effect of administrative hierarchical relationships on the co-occurrence of toponyms in turn.

Some other co-occurrence networks have been investigated, especially in language processing. Liang et al. [15–17] constructed character (word) co-occurrence networks for Chinese and English poems and analyzed their overall structural characteristics from a linguistic perspective. Compared with these networks, the toponym co-occurrence networks have more similar indicators, such as small world and disassortative traits. As language elements, toponyms also exhibit common linguistic features. However, there are still some major differences, including $\langle k \rangle$, γ and $C\%$. The $\langle k \rangle$ of the character (word) co-occurrence network is 2.39–6.01, γ is 1.98–2.83 and $C\%$ is 2.1–11.2. The toponym co-occurrence networks have far larger $\langle k \rangle$ because they are denser than poetry co-occurrence networks, and the average vertex degree is greater. There are more low-degree nodes and fewer high-degree nodes in toponym co-occurrence networks, so they have larger clustering coefficients and smaller degree distribution exponents.

6. Conclusion

This paper proposes a model of a toponym co-occurrence network, which is a new way to organize co-occurrence toponyms and consequently provides a promising approach for extracting valuable geographical information from massive web texts. The toponym co-occurrence networks are constructed in our experiment, while complex network analysis is applied to investigate their structure properties. The toponym co-occurrence network is found to have a small average shortest path length and a large clustering coefficient, indicating a strong local cluster feature and tendency toward internal aggregation, which has the small-world characteristic. The distribution degree of toponym vertices in the networks satisfies the power law, which means that these networks are scale-free. Meanwhile, the networks are disassortative, with assortative coefficients smaller than zero. Furthermore, these networks have a similar exponent μ, which denotes that co-occurrence networks constructed from the same topic corpus have a similar disassortative feature. The above structure properties are all independent of the network scale. The frequency of toponym co-occurrence is weakly negatively correlated with geographic distance but more strongly negatively correlated with administrative hierarchical distance. A link-based method is proposed to explore the core toponyms in hot events based on the transformation relationships of the toponyms.

The toponym co-occurrence network can contribute greatly to finding useful information in web texts and could even improve the accuracy of geographical information retrieval. Once the geographical information is extracted from massive web pages, the relationships of different toponyms are defined, which also provides a new opportunity to investigate the relatedness of web documents. Specific co-occurring toponyms cohere with different subjects, so certain toponyms can be found within a given topic. This information can be exploited for the discovery of related topics and vertical searches for geographic information retrieval.

Acknowledgment

This research was supported by the National Natural Science Foundation of China (Grant No. 41271385).

References

[1] C. Becker, C. Bizer, Exploring the geospatial semantic web with dbpedia mobile, Web Semant.: Sci. Serv. Agents World Wide Web 7 (4) (2009) 278–286.
[2] M.F. Goodchild, Citizens as sensors: the world of volunteered geography, GeoJournal 69 (4) (2007) 211–221.
[3] Y. Gao, S. Gao, R.Q. Li, et al., A semantic geographical knowledge wiki system mashed up with google maps, Sci. China Technol. Sci. 53 (1) (2010) 52–60.
[4] Y. Liu, Y. Yuan, D. Xiao, et al., A point-set-based approximation for areal objects: A case study of representing localities, Comput. Environ. Urban Syst. 34 (1) (2010) 28–39.
[5] Ž. Jakir, Ž. Hećimović, Z. Štefan, Place Names Ontologies, in: Advances in Cartography and GIScience, vol. 1, Springer Berlin, Heidelberg, 2011, pp. 331–348.
[6] F. Jiang, W. Wang, Research on Chinese toponym ontology mode, in: Information Science and Management Engineering (ISME), 2010 International Conference of, Vol. 1, IEEE, 2010, pp. 154–157.
[7] S. Overell, J. Magalhaes, S. Rüger, Place disambiguation with co-occurrence models, in: CLEF 2006 Workshop, Working notes. 2006.
[8] S. Overell, S. Rüger, Using co-occurrence models for placename disambiguation, Int. J. Geogr. Inf. Sci. 22 (3) (2008) 265–287.
[9] S.E. Overell, S. Rüger, Geographic co-occurrence as a tool for gir, in: Proceedings of the 4th ACM Workshop on Geographical Information Retrieval, ACM, 2007, pp. 71–76.
[10] D. Buscaldi, Approaches to disambiguating toponyms, SIGSPATIAL Spec. 3 (2) (2011) 16–19.
[11] Y. Liu, F. Wang, C. Kang, et al., Analyzing relatedness by toponym Co-occurrences on web pages, Trans. GIS 18 (1) (2014) 89–107.
[12] B. Roark, E. Charniak, Noun-phrase co-occurrence statistics for semiautomatic semantic lexicon construction, in: Proceedings of the 17th International Conference on Computational Linguistics-Vol. 2, Association for Computational Linguistics, 1998, pp. 1110–1116.
[13] W. Liang, Y. Shi, K.T. Chi, et al., Comparison of co-occurrence networks of the Chinese and English languages, Physica A 388 (23) (2009) 4901–4909.
[14] W. Liang, Y. Wang, Y. Shi, et al., Co-occurrence network analysis of modern Chinese poems, Physica A 420 (2015) 284–293.
[15] W. Liang, Y. Wang, Y. Shi, et al., Co-occurrence network analysis of modern Chinese poems, Physica A 420 (2015) 284–293.
[16] H.D. White, B.C. Griffith, Author cocitation: A literature measure of intellectual structure, J. Am. Soc. Inf. Sci. 32 (3) (1981) 163–171.
[17] L. Leydesdorff, C. Wagner, H.W. Park, et al. International collaboration in science: The global map and the network, 2013. ArXiv Preprint arXiv:1301.0801.
[18] J.P. Qiu, K. Dong, H.Q. Yu, Comparative study on structure and correlation among author co-occurrence networks in bibliometrics, Scientometrics 101 (2) (2014) 1345–1360.
[19] H. Small, Co-citation in the scientific literature: A new measure of the relationship between two documents, J. Amer. Soc. Inf. Sci. 24 (4) (1973) 265–269.
[20] R. Mihalcea, P. Tarau, TextRank: Bringing Order into Texts, Association for Computational Linguistics, 2004.
[21] A. Heydon, M. Najork, Mercator: A scalable, extensible web crawler, World Wide Web 2 (4) (1999) 219–229.
[22] W. Che, Z. Li, T. Liu, Ltp: A Chinese Language Technology Platform. Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations, Association for Computational Linguistics, 2010, pp. 13–16.
[23] D.A. Schult, P. Swart, Exploring network structure, dynamics, and function using NetworkX. in: Proceedings of the 7th Python in Science Conferences, SciPy 2008. Vol. 2008, 2008, pp. 11–16.
[24] M. Rubinov, O. Sporns, Complex network measures of brain connectivity: uses and interpretations, Neuroimage 52 (3) (2010) 1059–1069.
[25] J. Scott, Social network analysis, Sociology 22 (1) (1988) 109–127.
[26] S.P. Borgatti, Centrality and network flow, Soc. Networks 27 (1) (2005) 55–71.
[27] S.N. Dorogovtsev, J.F.F. Mendes, A.N. Samukhin, Size-dependent degree distribution of a scale-free growing network, Phys. Rev. E 63 (6) (2001) 062101.
[28] S.N. Dorogovtsev, A.V. Gotsev, J.F.F. Mendes, Pseudofractal scale-free web, Phys. Rev. E 65 (2002) 066122.
[29] A.-L. Barabási, E. Ravasz, T. Vicsek, Determinic scale-free networks, Physica A 229 (2001) 559–564.
[30] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509–512.
[31] J.S. Andrade, H.J. Herrmann, R.F.S. Andrade, L.R. da Silva, Errattu: Apollonian networks, Phys. Rev. Lett. 102 (2009) 079901.
[32] E.W. Dijkstra, A note on two problems in connexion with graphs, Numer. Math. 1 (1) (1959) 269–271.
[33] D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, Nature 393 (6684) (1998) 440–442.
[34] P. Erd6s, A. Rényi, On the evolution of random graphs, Publ. Math. Inst. Hung. Acad. Sci. 5 (1960) 17–61.
[35] R. Pastor-Satorras, A. Vázquez, A. Vespignani, Dynamical and correlation properties of the Internet, Phys. Rev. Lett. 87 (25) (2001) 258701.
[36] M.E.J. Newman, Assortative mixing in networks, Phys. Rev. Lett. 89 (20) (2002) 208701.
[37] C. Wu, L. Shen, X. Wang, A new method of using contextual information to infer the semantic orientations of context dependent opinions, in: International Conference on Artificial Intelligence and Computational Intelligence, 2009, Vol. 4, AICI'09, IEEE, 2009, pp. 274–278.
[38] S. Brin, L. Page, The anatomy of a large-scale hypertextual Web search engine, Comput. Netw. 30 (1998) 107–117.
[39] P. Boldi, M. Santini, S. Vigna, PageRank as a function of the damping factor, in: Proceedings of the 14th International Conference on World Wide Web, ACM, 2005, pp. 557–566.
[40] M. Henzinger, Link analysis in web information retrieval, IEEE Data Eng. Bull. (2000) 3–8.
[41] W.R. Tobler, A computer movie simulating urban growth in the Detroit region, Econ. Geogr. 46 (1970) 234–240.