ABSTRACT: This paper used China’s urban clusters as case studies, and a model of the information network was constructed using Baidu index search data. In this model, the urban clusters were considered nodes, and the search intensity between them were regarded as edges. Based on complex network theory and city flow theory, the structure of geographical space was visualized in the form of a network. Analysis results show that, first, the network density of the core contact is only at 3.62%, and 69.2% of the connection strengths between urban clusters are relatively weak on a national scale. Next, the abilities of absorbing information are greater than those of giving off for more developed urban clusters. Last, compared with the ability of giving off information, the variability of absorption is more obvious for all 24 urban clusters. Based on these conclusions, we analyzed spatial influences from the perspective of urban flows and present strategies with space optimization.

1. Introduction

Connections between cities are constantly being reconstructed given the effects of globalization and information traffic. With an expanding economy, the urban cluster has become the most dynamic hub at either the global, national, or regional levels. Meanwhile, urban networks have become a more effective concept of spatial structure in related studies. Given that select social-media sites that support submission and sharing of geo-tagged information have achieved considerable commercial success, the age of big data is said to have already dawned[1–3]. With the aid of data sources from the websites, such as Google, Facebook, and Twitter, constructing a more precise model of geographical space has become feasible[4–5].

Recently, in research on geographical space, big data mining and its application primarily include the following aspects: (1) visualizing social space structure with social media sites[6–7], (2) analyzing urban vitality with population statistics[8], (3) optimizing the balance of career and habitation with geo-tagged information[9], (4) evaluating services usage with mobile location[10], (5) scouting locations with data for customer behavior[11], (6) building an efficient logistics system with e-commerce data[12], and (7) improving government efficiency with urban inspection data[13–15].

From the perspective of complex network theory[16–17], many systems were abstracted into networks, such as the Internet[18], biological nets[19], food webs[20], and other areas. Based on city flow theory, cities and regions were also seen as spatial networks shaped by flows. In
related studies, scholars in China and abroad analyzed the industry network[21], subway network[22], street network[23], and population moving network[24], among others. The present paper tries to combine the two theories and construct a geographical space model based on statistical data from location-based service (LBS).

Unlike traditional studies, we directly obtained the data of contact strengths between urban clusters from websites instead of the gravity model. In this research, the searched data for 193 cities and 24 urban clusters in China (see Figure 1) were used to understand how information networks are shaping new geographical spaces on a national scale. This study has three aims. The first is to analyze network density, which ranges in three different thresholds. Next, we plan to determine the distribution of vertex strengths, including both in strength and out. Last, we intend to investigate the structure entropy of the network to describe the balance of space.

2. Method Descriptions

The networks were built based on graph theory. We used graph \( G = (V, E) \) to express the complex networks of urban clusters, where \( V \) is a set of nodes and \( E \) is a set of edges. The urban clusters were regarded as the nodes, as depicted by \( V = \{ v_1, v_2, \ldots, v_n \} \), where \( n \) is the number of urban clusters. The search intensity between the clusters were regarded as the edges, as depicted by \( E = \{ e_1, e_2, \ldots, e_m \} \), where \( m \) is the number of connections.

Based on space model, we simulated the networks of China’s urban clusters and researched their properties following the four steps below:

Step 1: Obtaining data of contact strengths between urban clusters from websites and transforming original statistical data into standard data.

Step 2: Visualization of the spatial network structure with coding standard data, which represent connection strengths between urban clusters.

Step 3: Analysis of the properties of network structure with the models of the complex network, including both nodes and edges.

Step 4: Determining distribution of geographical space based on network structure properties.

3. Data Processing

Baidu is a Chinese Internet giant that operates a search engine. Baidu Index, a type of search data similar to Google Trends, is one of Baidu’s most successful new projects in terms of usage. It lets customers access correlation data between cities via Baidu LBS. As shown in Figure 2, we can obtain correlation data from city \( a \) to \( b \) by setting the name of \( a \) as the keyword and \( b \) as the search area in the Baidu index, as depicted by \( w_{ab} \). Based on this approach, data for information contact strengths between each pair of 193 cities in China from January to February this year can be obtained. We then counted the original values of the information contact strength from urban clusters \( A \) to \( B \) via the method below:

\[
W_{AB} = \sum W_{\text{ab}}
\]

where \( W_{AB} \) is the correlation data from urban cluster \( A \) to \( B \), \( A = \{ a \}, B = \{ b \} \) (see Figure 2). The original statistical data were then transformed into standard data via the following method:
where $\hat{W}_{AB}$ is the standard correlation data from urban cluster $A$ to $B$, and $w_{\text{max}}$ is the maximum value of all pairs of clusters. Given the data for contact strengths between urban clusters, a relative adjacency matrix can be constructed, which can fully describe the network structure via the method of graph theory.

\[ \hat{W}_{AB} = \frac{w_{AB}}{w_{\text{max}}} \times 100 \]  

(2)

4. Models Visualizing

Other studies have visualized a geographical network space with Ucinet, Netdraw, Gephi, R, or other software.

In the present research, we attempted to use Echart, a language based on JavaScript, to describe the map type, latitudes, and longitudes of core cities and directions and strengths of relationships between urban clusters. As shown in Figure 3(a), the graph is a directed and weighted network, which consists of 24 nodes and 552 edges.

It provides intuitive, accurate, and highly customized data visualization.

5. Network Analysis

5.1 Network Density

In static network analysis, the distribution of edges is usually described by network density. The network density is defined as follows:

\[ \rho = \frac{M}{N(N-1)} \]  

(3)

where $\rho$ is the network density, $M$ is the number of connections, $N$ is the number of nodes. As shown in Figure 3, we divided the complete graph into three threshold ranges, which are defined as core connection network, medium connection network, and weak connection network. Based on the calculation results of Formula 3, the network density of the core connection network is only 3.62%, and most of these connections originate from the Yangtze River Delta, Jingjinji, Zhusanjiaozhou, and Chengyu urban cluster. The network density of the medium connection network is 27.17%, whose connections are almost located in the more dynamic urban clusters in southeast China. Finally, the density of the weak connection network is
69.20%, which indicates a low level of contact between most urban clusters.

5.2 Vertex Strength
Vertex strength, a concept corresponding to edge weight, is defined as follows:

\[ S_i = \sum_{j \in N_i} w_{ij} \]  \hspace{1cm} (4)

where \( S_i \) is the vertex strength of the node \( v_i \), \( N_i \) is a set of nodes adjacent to \( v_i \), and \( w_{ij} \) is the edge weight between \( v_i \) and \( v_j \). In a directed network, the vertex strength can also be expressed as

\[ S_i = S_i^{in} + S_i^{out} \]  \hspace{1cm} (5)

where \( S_i^{in} \) is the in strength, which represents the combined edge weights from adjacent nodes to node \( v_i \), and \( S_i^{out} \) is the out strength, which represents the combined edge weights from node \( v_i \) to adjacent nodes.

Based on the results shown in Figure 4, the following results were obtained: (1) the Yangtze River Delta, Zhusanjiaozhou, Jingjinji, Chengyu, and Shandongbandao urban clusters are the regions that have the maximum values of vertex strengths, (2) the Lhasa, Yinanchuan, Tianshanbei, and Qianzhong urban clusters have the minimum values of vertex strengths, (3) the in strengths are usually larger than the out strengths for developed regions, thereby indicating that these regions tend to receive rather than give information.

5.3 Space Balance
Space balance can be described by the structure entropy of a network. Assuming that \( S_i \) is the vertex strength of node \( v_i \), its importance can be depicted by:

\[ I_i = \frac{S_i}{\sum_{j=1}^{N} S_j} \]  \hspace{1cm} (6)

where \( N \) is the number of nodes, the nodes whose \( S = 0 \) is not considered. The structure entropy of the network can be depicted by:

\[ E = -\sum_{i=1}^{N} I_i \cdot \ln I_i \]  \hspace{1cm} (7)

To remove the effect of node number, the entropy is transformed into standard data via the following method:

\[ \frac{E - E_{\min}}{E_{\max} - E_{\min}} = \frac{-2 \sum_{i=1}^{N} I_i \cdot \ln I_i - \ln 4(N-1)}{2 \ln N - \ln 4(N-1)} \]  \hspace{1cm} (8)

where \( E \) is the standard entropy, \( E \) is the absolute entropy, \( E_{\min} \) is the theoretical minimum value of \( E \), \( E_{\max} \) is the theoretical maximum of \( E \), and \( N \) is the number of nodes.

Obviously, \( E \) ranges from 0 to 1. These formulas can be used to obtain the logarithmic distribution of in strengths, out strengths, and vertex strengths (see Figure 5). Furthermore, the corresponding standard entropies are 0.663, 0.805, and 0.745, respectively. These results mean that the variability of absorption is more obvious for all 24 urban clusters compared with the ability of giving off information.

6. Influence Factors Analysis
Geographical space is shaped by various city flows. To determine the correlations between network structure in the model and the spatial flows, we collect data for...
Figure 5. Distributions of the vertex strengths

![Distributions of the vertex strengths](image)

<table>
<thead>
<tr>
<th>Urban cluster</th>
<th>Webpage (million)</th>
<th>Institute</th>
<th>Train</th>
<th>GDP (billion yuan)</th>
<th>Population (million)</th>
<th>AQI</th>
</tr>
</thead>
<tbody>
<tr>
<td>CY</td>
<td>16.920</td>
<td>2395</td>
<td>664</td>
<td>3325.8</td>
<td>96.55</td>
<td>94</td>
</tr>
<tr>
<td>CZ</td>
<td>6.959</td>
<td>819</td>
<td>869</td>
<td>1766.3</td>
<td>40.47</td>
<td>85</td>
</tr>
<tr>
<td>WH</td>
<td>6.478</td>
<td>1304</td>
<td>983</td>
<td>1344.1</td>
<td>36.83</td>
<td>75</td>
</tr>
<tr>
<td>JH</td>
<td>8.692</td>
<td>795</td>
<td>670</td>
<td>969.9</td>
<td>29.86</td>
<td>58</td>
</tr>
<tr>
<td>HP</td>
<td>4.845</td>
<td>661</td>
<td>700</td>
<td>9883.8</td>
<td>128.73</td>
<td>79</td>
</tr>
<tr>
<td>CS</td>
<td>55.314</td>
<td>7528</td>
<td>4004</td>
<td>1734.9</td>
<td>23.38</td>
<td>99</td>
</tr>
<tr>
<td>LB</td>
<td>2.435</td>
<td>580</td>
<td>162</td>
<td>349.8</td>
<td>13.76</td>
<td>105</td>
</tr>
<tr>
<td>YC</td>
<td>1.295</td>
<td>209</td>
<td>51</td>
<td>201.5</td>
<td>5.39</td>
<td>101</td>
</tr>
<tr>
<td>ZY</td>
<td>10.522</td>
<td>1376</td>
<td>707</td>
<td>1734.9</td>
<td>41.95</td>
<td>140</td>
</tr>
<tr>
<td>GZ</td>
<td>4.783</td>
<td>1236</td>
<td>421</td>
<td>874.1</td>
<td>23.38</td>
<td>90</td>
</tr>
<tr>
<td>HB</td>
<td>4.296</td>
<td>364</td>
<td>274</td>
<td>1186.1</td>
<td>12.27</td>
<td>93</td>
</tr>
<tr>
<td>TY</td>
<td>4.356</td>
<td>634</td>
<td>387</td>
<td>830.1</td>
<td>23.53</td>
<td>79</td>
</tr>
<tr>
<td>JJ</td>
<td>56.798</td>
<td>5964</td>
<td>2045</td>
<td>5201.8</td>
<td>86.84</td>
<td>124</td>
</tr>
<tr>
<td>SD</td>
<td>12.209</td>
<td>2612</td>
<td>792</td>
<td>3164.8</td>
<td>44.30</td>
<td>100</td>
</tr>
<tr>
<td>ZS</td>
<td>96.305</td>
<td>2635</td>
<td>2383</td>
<td>6600.8</td>
<td>64.56</td>
<td>52</td>
</tr>
<tr>
<td>PX</td>
<td>9.215</td>
<td>1281</td>
<td>1008</td>
<td>1971.3</td>
<td>37.48</td>
<td>48</td>
</tr>
<tr>
<td>NN</td>
<td>3.710</td>
<td>613</td>
<td>279</td>
<td>590.1</td>
<td>19.98</td>
<td>56</td>
</tr>
<tr>
<td>QH</td>
<td>3.220</td>
<td>378</td>
<td>178</td>
<td>583.3</td>
<td>18.41</td>
<td>41</td>
</tr>
<tr>
<td>DZ</td>
<td>2.472</td>
<td>728</td>
<td>149</td>
<td>598.1</td>
<td>17.52</td>
<td>49</td>
</tr>
<tr>
<td>QZ</td>
<td>2.468</td>
<td>434</td>
<td>144</td>
<td>447.6</td>
<td>19.55</td>
<td>48</td>
</tr>
<tr>
<td>HC</td>
<td>9.813</td>
<td>1758</td>
<td>758</td>
<td>1821.9</td>
<td>33.89</td>
<td>76</td>
</tr>
<tr>
<td>LZ</td>
<td>7.668</td>
<td>1469</td>
<td>791</td>
<td>2302.5</td>
<td>30.61</td>
<td>88</td>
</tr>
<tr>
<td>TS</td>
<td>1.760</td>
<td>380</td>
<td>92</td>
<td>442.2</td>
<td>7.41</td>
<td>67</td>
</tr>
<tr>
<td>LS</td>
<td>0.543</td>
<td>52</td>
<td>18</td>
<td>51.3</td>
<td>1.97</td>
<td>46</td>
</tr>
</tbody>
</table>

Table 1. Spatial factors of China’s urban clusters

<table>
<thead>
<tr>
<th>Webpage</th>
<th>Institute</th>
<th>Train</th>
<th>GDP</th>
<th>Population</th>
<th>AQI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{wi}$</td>
<td>0.861</td>
<td>0.929</td>
<td>0.972</td>
<td>0.984</td>
<td>0.920</td>
</tr>
<tr>
<td>$S_{oi}$</td>
<td>0.807</td>
<td>0.864</td>
<td>0.940</td>
<td>0.958</td>
<td>0.933</td>
</tr>
<tr>
<td>$S_{i}$</td>
<td>0.846</td>
<td>0.909</td>
<td>0.967</td>
<td>0.982</td>
<td>0.934</td>
</tr>
</tbody>
</table>

Table 2. The correlation analysis of spatial factors

Note: (a) Webpage means the number of webpages in the urban clusters and represents the information factor. The data are derived from Baidu search in February 2015. (b) Institute means the number of research institutes in the urban clusters and represents the innovation factor. The data are derived from Baidu LBS in February 2015. (c) Train means the number of trains in the urban clusters and represents the transportation factor. The data are derived from service center of China railway in February 2015. (d) GDP represents the economy factor. The data are derived from Chinese city yearbook in 2012. (e) Population represents the manpower factor. The data are derived from Chinese city yearbook in 2012. (f) AQI means air-quality index in urban clusters and represents the environment factor. The data are derived from an air quality website in February 2015.
information, innovation, transportation, economy, population, and environment, as shown in Table 1. We then analyzed the correlation of the data collected using SPSS software, and the results are shown in Table 2. All of the spatial factors have a significant correlation with the strengths of nodes, except for AQI. In addition, economy and transportation are the two most important factors that affect China’s geographical space. With relatively low impact levels, more attention should be paid to innovation and information.

6. Conclusions

With the rapid development of big data, a new perspective was proposed to modify conventional geographical analyses. Based on the search data provided by Baidu LBS, this study constructed a model of spatial network in China. The data of the geographical space can be visualized by analyzing the relationships between urban clusters. To understand the structural features of space, we analyzed the network density, vertex strength, and space balance. The results indicate that, first, network density of the core contact is at a low level, and most of the connection strengths between urban clusters are relatively weak. Next, for developed urban clusters, their abilities of absorbing information are greater than those of giving them off. Last, the distribution of absorbing information is more variable compared with that of giving them off. Moreover, no correlation was observed between environmental factors and social geography. Other factors, such as economy, transportation, population, innovation, and information, have a significant correlation with the strengths of nodes.

Based on these findings, optimization strategies can be proposed in two aspects: the first is to enhance the vertex strength of Chengyu, Sandongbandao, Liaozhongnan and other urban clusters, thereby increasing the network density of core contact and improving the spatial mobility on a national scale. Second, for less developed regions, their abilities of absorbing information are greater than those of giving them off. Moreover, no correlation was observed between environmental factors and social geography. Other factors, such as economy, transportation, population, innovation, and information, have a significant correlation with the strengths of nodes.

This paper avoids the shortcomings of the lack of direct data and undirected relationships between regions. However, this paper ignores the effects of time, which can be improved in future studies to achieve a more comprehensive and precise method for understanding the spatio-temporal changes in geography.

Acknowledgement

This work was supported by Chinese National Natural Science Foundation (No. 51478470), Soft Science Research Project of Hunan Province (No. KY201305) and Graduate Innovative Research Project of Hunan Province (No. CX2011B100).

References


