Analysis of Current Utilization Status of Urban Government Open Data in China from Thematic and Regional Perspectives: Postprint

Authors: Duan Yaoqing, Xueting Qiu, He Siqi

Date: 2023-08-27T00:00:00+00:00

Abstract

[Purpose/Significance] To analyze the utilization status of urban government open data in China from both thematic and regional perspectives at the data level, investigate the linear relationship between attention level and utilization level of open data, and thereby enhance the usage efficiency of government open data in China. [Method/Process] Using six cities—Harbin, Jinan, Shanghai, Wuhan, Guangzhou, and Guiyang—as case studies, this research conducts statistical comparative analysis, cluster analysis, and regression analysis on multiple indicators including view rate and download rate of open data, to reveal the utilization status of urban open data in China from thematic and regional viewpoints, and to examine the linear relationship between view rate and download rate. [Results/Conclusion] The utilization of urban government open data in China demonstrates the following characteristics: overall, a weak correlation exists between the view rate and download rate of open data. From a thematic perspective, topics related to social livelihood—including education and technology, people's livelihood services, and economic and business affairs—exhibit high utilization levels, with a positive correlation between their view rate and download rate; the overall characteristics of open data are inconsistent with some of its specific features. From a regional perspective, Jinan and Shanghai display a positive correlation between view rate and download rate, with Shanghai ranking first in open data utilization and Jinan ranking last; Guiyang shows relatively high view and download rates for open data, but with a weak correlation between them; the overall characteristics of open data are largely consistent with some of its specific features.

Full Text

Preamble

ChinaXiv Partner Journal
Volume 62, Issue 20, October 2018

Analysis of the Utilization Status of Urban Government Open Data in China from Thematic and Regional Perspectives

Duan Yaoqing, Qiu Xueting, He Siqi
School of Information Management, Central China Normal University, Wuhan 430079

Abstract

[**Purpose/Significance**] This paper analyzes the utilization status of urban government open data in China from both thematic and regional perspectives at the data level, and explores the linear relationship between the degree of attention and utilization of open data to improve the efficiency of government open data usage. [**Method/Process**] Taking Harbin, Jinan, Shanghai, Wuhan, Guangzhou, and Guiyang as examples, the study conducts statistical comparative analysis, cluster analysis, and regression analysis on multiple indicators such as browsing rate and download rate of open data to reveal the utilization status of urban open data in China from thematic and regional perspectives, and examines the linear relationship between browsing rate and download rate. [**Result/Conclusion**] The utilization of urban government open data in China exhibits the following characteristics: overall, there is a weak correlation between browsing rate and download rate of open data. From a thematic perspective, topics related to social livelihood such as education and technology, people's livelihood services, and economic and industrial commerce show high utilization, with positive correlation between browsing rate and download rate; the overall characteristics of open data are inconsistent with some individual features. From a regional perspective, Jinan and Shanghai show positive correlation between browsing rate and download rate, with Shanghai ranking first in open data utilization and Jinan ranking last; Guiyang has high browsing and download rates for open data, but they are weakly correlated; the overall characteristics of open data are generally consistent with some features.

Classification Number: D63 G203
Keywords: government open data; correlation analysis; utilization

1. Introduction

Open government data refers to data produced by government, government-commissioned, and government-controlled entities that can be freely used, reused, and redistributed by anyone [1]. This movement first emerged in the United States in 2009 and has received significant attention from governments and academia worldwide. According to the "Open Data Barometer" global

report released in April 2016, 114 countries have joined this initiative [2]. In China, the launch of the "Shanghai Municipal Government Data Service Network" pilot in 2012 marked the beginning of open data. By November 2017, 23 provinces, municipalities, or district governments had established local government data open platforms (excluding Hong Kong, Macao, and Taiwan). The utilization of government open data refers to the extent to which open data resources meet people's needs and usage, with its essence being the effective allocation and use of resources. Current research on the utilization status of government open data both domestically and internationally is relatively macroscopic and has achieved certain results.

Foreign academic research on government open data utilization focuses on two aspects: first, research on utilization barriers of open data. Initially, it was recognized that the level of government data openness directly affects the service effectiveness of government data [3], while simultaneously identifying implementation barriers and usage barriers in open data [4]. Subsequently, barriers such as accessibility, usability, and interactivity in the commercial utilization of open data were discovered [5], leading to the construction of evaluation models to identify user usage and satisfaction [6], and further analysis of barrier mechanisms. Second, research on the utilization value of open data. Theoretically, it has been recognized from evaluating the motivations of local governments to open data [7] that the potential value of government open data should be activated through effective utilization models [8]. In practice, relevant conceptual models [9] and visualization methods [10] have been applied to mine, understand, and communicate the value of open data.

Domestic academic research on open data utilization focuses on three aspects: first, evaluation research on open data utilization. For government data open platforms, corresponding evaluation frameworks, indicators, and methods have been proposed [11], and existing government data open platforms in China have been evaluated from the perspective of user utilization [12]. Furthermore, several platforms in Guangdong, Beijing, Shanghai, and other places have been divided into three levels according to service performance [13]. For open data, indicators such as dataset visits, downloads, application status [14], the ratio of downloads to views, and average views [15] have been selected to measure the utilization effectiveness of government open data. Second, research on guarantee mechanisms for open data. The efficient utilization of government open data cannot be achieved without adequate guarantees in law, technology, data sharing, and user participation [16], while the integrity, accuracy, timeliness of open data and national policies also affect data utilization [17]. Third, research on utilization methods of open data. Macroscopically, factors affecting the utilization of government open data have been explored [18], while microscopically, utilization methods such as API interfaces and APP program development have been incorporated into the scope of data utilization [19]. By reviewing domestic and foreign government data open utilization methods, the viewpoint has been proposed that the improvement of open data value is proportional to data utilization rate [20].

In summary, domestic and foreign research focuses on exploring the construction of data open platforms from a macro perspective, with few studies analyzing the utilization status of open data from a micro perspective. The ultimate purpose of government open data is to promote its use and development. To help governments maximize public data needs and understand the gaps in data utilization among Chinese urban governments, this study takes the data itself as the entry point. From thematic and regional perspectives respectively, it uses indicators such as data browsing rate and download rate to compare and analyze the utilization degree of open data, and uses cluster analysis and correlation analysis to grasp the utilization status of open data from different themes and different city governments, thereby promoting the work of government open data in China.

2. Sample Selection and Data Collection

2.1 Sample Selection

Since most Chinese government data open platforms use "data.gov.cn" as their domain name, this paper searches using this domain. As of November 29, 2017, 23 cities in China had established open data platforms. From the existing platforms, open data mainly focuses on various aspects of social life such as economy, transportation, education, and environmental protection. However, there are significant differences among cities in the number and names of data themes, with similar data having different names on different platforms and data from different themes belonging to the same category. Therefore, this paper first reviews the theme classification of the existing 23 data open platforms. To improve research focus and efficiency, it organizes and summarizes 297 resource themes from all datasets on existing platforms into 19 major categories such as economic and industrial commerce, fiscal and financial affairs, and statistics the distribution of sub-themes in each major category. The specific distribution is shown in Figure 1 [Figure 1: see original paper].

2.1.1 Thematic Perspective Sample Selection From the statistical analysis of Chinese government open data theme classification, the distribution of existing open data themes shows a state of concentration and dispersion. In terms of quantity, the "culture, sports, and leisure" category contains 29 sub-themes, the most among all major categories, while "religious belief" and "judicial services" contain only 5 sub-themes each, the fewest. In terms of open fields, current local government open data focuses on livelihood areas related to education, employment, and healthcare (culture, sports, and leisure; social security and employment; transportation; education and technology; medical and health), ecological fields (energy and environment), and social governance fields (government agencies and social organizations, public security). The theme classification shows that local governments have both commonalities and certain particularities in classifying open data.

Based on the common characteristics of theme classification (referring to certain

themes appearing simultaneously on platforms) and the distribution characteristics of regional samples (see 2.1.2 for regional sample selection) and their coverage on each platform, after removing the "other" category, 10 major categories are selected as preliminary research samples from the thematic perspective: "culture, sports, and leisure," "social security and employment," "economic and industrial commerce," "energy and environment," "transportation," "people's livelihood services," "education and technology," "government agencies and social organizations," and "public security."

2.1.2 Regional Perspective Sample Selection Cities with open data have broad representativeness in administrative level and geographical region. Administratively, these 23 cities include provincial-level cities, sub-provincial cities, prefecture-level cities, and county-level cities. Geographically, they are distributed across multiple municipalities or provinces such as Beijing, Shanghai, Hubei, and Guangdong.

Due to differences in administrative levels, the focus and progress of data openness vary significantly among platforms, making comprehensive sample analysis unsuitable. Therefore, after multiple discussions, this paper selects six cities as research samples: Jinan, Shanghai, Wuhan, Guangzhou, Harbin, and Guiyang. The comparability of these six cities is reflected in two aspects: first, except for Wuhan not covering "social security and employment" and Guiyang not covering "people's livelihood services," the remaining eight major themes correspond to the six research samples; second, these six samples are all municipalities or provincial capitals in terms of administrative level and belong to different regions geographically, making them representative and meaningful as regional samples. Combining the determination of research samples from both thematic and regional perspectives, this paper selects 63 sub-themes from the six platforms, categorizes them into 10 major themes, and finalizes the research samples as shown in Table 1 .

2.2 Data Collection

According to statistics, the six government data open platforms have a total of 86 sub-themes, with this study covering 63 sub-themes. The Octopus data collector and manual observation methods are used to capture relevant information on datasets from the 63 sub-themes across the six platforms for subsequent experiments. All data collection was completed by December 12, 2017.

2.3 Parameter Determination for Open Data Utilization Status

Current research mostly measures the utilization effectiveness of open data from macro perspectives such as views and downloads. Scholars believe that data views and downloads affect user attention and utilization effectiveness [21], and compare the utilization effectiveness of open data in Beijing and Shanghai by calculating the ratio of downloads to views [22]. Some also select visits and downloads to measure the utilization status of some local government data [13].

However, indicators such as browsing rate and download rate better reflect the proportion of local data in overall data, thereby reflecting user attention and utilization degree toward certain themes or regional open data. Therefore, this study selects the total amount of open data as NC, total browsing volume as BC, and total download volume as DC. The specific formulas are shown in (4)-(6):

where i represents the 6 cities (i=1,2,…,6), nci represents the number of open datasets in city i, bci represents the browsing volume of open data in city i, and dci represents the download volume of open data in city i.

Based on these formulas, this study first statistically analyzes the browsing rate and download rate of open data from thematic and regional perspectives to reveal user attention and utilization degree, and then conducts regression analysis on browsing rate and download rate to identify their correlation and help realize the value and purpose of government data. Other calculation formulas and methods are shown in Table 2 .

3. Data Analysis

3.1 Thematic Perspective on Government Open Data Utilization Status

To reveal the utilization status of open data under different themes, let the total number of open datasets for all 10 major theme categories in Table 1 be NT, total browsing volume be BT, and total download volume be DT. The specific formulas are shown in (1)-(3):

where each major theme category includes several sub-themes. i represents the 10 major theme categories (i=1,2,…,10); nti represents the number of open datasets in theme category i; bti represents the browsing volume of open data in theme category i; dti represents the download volume of open data in theme category i.

Similarly, let the total number of open datasets for all themes in the 6 cities in Table 1 be NC, total browsing volume be BC, and total download volume be DC. The specific formulas are shown in (4)-(6):

where i represents the 6 cities (i=1,2,…,6), nci represents the number of open datasets in city i, bci represents the browsing volume of open data in city i, and dci represents the download volume of open data in city i.

Based on these formulas, this study first statistically analyzes the browsing rate and download rate of open data from thematic and regional perspectives to reveal user attention and utilization degree, and then conducts regression analysis on browsing rate and download rate to identify their correlation and help realize the value and purpose of government data. Other calculation formulas and methods are shown in Table 2 .

3.1.1 Browsing Rate of Open Data from Thematic Perspective Browsing rate can intuitively reflect user attention to data on a certain theme. First, R is used to plot line charts and scatter plots of browsing rates for open data in each major theme category, as shown in Figure 4 [Figure 4: see original paper]. The X-axis represents the 10 major theme categories, i.e., ti (i=1,2,…,10). Themes 1-10 represent: culture, sports, and leisure; economic and industrial commerce; transportation; medical and health; government agencies and social organizations; social security and employment; energy and environment; people's livelihood services; education and technology; and public security.

As shown in Figure 2, the top three themes in terms of browsing volume are economic and industrial commerce (937,657 times), transportation (420,002 times), and people's livelihood services (327,778 times). Among them, transportation data also has the highest average browsing rate per single sample, indicating that data on themes such as economic and industrial commerce and transportation receive significant user attention. However, Figure 3 shows that some themes have values significantly below the average for this indicator, such as government agencies and social organizations, whose average browsing rate is less than 300 times per entry, indicating differences in attention levels across themes. Figure 4 shows that the theme with the highest browsing rate, economic and industrial commerce (approximately 0.267), is 4.94 times that of the lowest, public security (approximately 0.054). Moreover, only economic and industrial commerce and transportation have browsing rates above the average (0.1) among these 10 categories.

Additionally, comparing the line chart and scatter plot of browsing rates in Figure 4 reveals that although economic and industrial commerce has the highest browsing rate, its scatter plot distribution shows that not every piece of data in this category has high browsing rates (most fall within 0-0.02). Conversely, public security data shows the opposite pattern: while its overall browsing rate is the lowest, the span of browsing rates for its open data is relatively large, mostly between 0.1-0.25, indicating that the overall and partial characteristics of open data across themes are not consistent.

3.1.2 Download Rate of Open Data from Thematic Perspective Download rate further elaborates on and deepens browsing rate, largely reflecting user utilization of certain data. Similar to browsing rate, Figure 5 [Figure 5: see original paper] shows line charts and scatter plots of download rates for open data in each major theme category, with themes 1-10 having the same meanings as in Figure 4 in section 3.1.2.

Figure 5 shows that users have deeper utilization of economic and industrial commerce, education and technology, social security and employment, and transportation data, reflected in two aspects. First, four themes have download rates above the average: economic and industrial commerce (approximately 0.223), education and technology (approximately 0.133), social security and employment (0.103), and transportation (approximately 0.1). Additionally, except for

social security and employment, the overall average browsing rates of the other three themes also exceed the corresponding averages, indirectly demonstrating the authenticity of user utilization of such data. Second, the sum of download rates for these four themes is approximately 0.56, accounting for more than half of the total downloads across all themes, indicating high utilization and suggesting that the utilization degree of other themes needs improvement.

Furthermore, combined with the download rate scatter plot, the distribution of download rates for other themes is uneven. Taking public security as an example, although its download rate scatter plot spans a wide range, the theme's download rate ranks at the bottom overall. Additionally, most themes, including economic and industrial commerce, have low overall average browsing rates for open data, with values mostly falling within 0-0.05.

3.1.3 Utilization Status of Open Data from Thematic Perspective To deeply reveal the utilization status and relationships of open data across themes, cluster analysis is performed on the 10 major theme categories. Cluster analysis refers to dividing data into different groups according to their own characteristics without predetermined grouping rules, minimizing internal differences within groups while maximizing differences between groups [23]. First, clustering indicators are selected. Since the average browsing rate (download rate) of single samples and overall samples change in the same direction and correspond one-to-one, with the former being the basis for calculating the latter, cluster analysis mainly refers to four indicators: open data browsing rate, download rate, and overall sample average browsing rate (download rate). Meanwhile, hierarchical clustering is used, with squared Euclidean distance for individual distances and Ward's method for inter-cluster distances. The final clustering results are shown in Figure 6 [Figure 6: see original paper].

To better determine the number of clusters, a scree plot for thematic perspective open data clustering is drawn, as shown in Figure 6(b). As clusters continue to merge and the number of categories decreases, the distances between categories increase rapidly, and the scree plot gradually flattens. Observing the scree plot shows that before clustering into 4 categories, the distances between categories are small, but after clustering into 4 categories, the distances become large. Therefore, 4 categories represent the "inflection point" of this scree plot, making 4 or 3 categories appropriate. After comprehensive consideration, this study ultimately clusters the 10 major themes into 4 categories, as shown in Figure 6(a).

Specifically, the first category includes "culture, sports, and leisure," "social security and employment," and "education and technology," which cluster into one category in just three steps with coefficients of 0.128 and 1.465. Social security and employment and education and technology data exceed the corresponding averages in three indicators including overall sample average browsing rate, with other indicators ranking relatively high and small variance and standard deviation. This indicates that livelihood-related data have high utilization

efficiency, and education and technology, social security and employment, and culture, sports, and leisure are among the daily issues users care about most, being closest to users and further demonstrating that user demand is the prerequisite for data utilization.

The second category is "economic and industrial commerce," which occupies an absolute advantage in four indicators including browsing rate and download rate, thus forming its own category. The high attention and utilization of this category are mainly contributed by Shanghai's open data. Such data mainly involve local economic construction and industrial and commercial trade information, covering economic, industrial, statistical, trade, consumption, and economic policy information. Users, especially enterprise users, pay more attention to and have greater demand for such data, resulting in stable and high indicators for economic and industrial commerce data.

The third category includes "transportation" and "medical and health" data, which cluster with medical and health in step 5 with a coefficient of 5.187. Although their overall browsing and download rates are not high, their overall sample average browsing and download rates rank at the top. Transportation is closely related to users' daily lives, and medical and health is a hot topic of concern for the whole society. The emergence of "Internet Plus" transportation and electronic medical care has greatly saved users' time and facilitated public life.

The fourth category includes "people's livelihood services," "energy and environment," "government agencies and social organizations," and "public security" data, whose four indicators are all negative and below average. The utilization of this category is relatively low, and users currently pay less attention to issues in social governance fields such as energy, environment, and public security, which is related to user awareness and the urgency of demand.

3.2 Regional Perspective on Government Open Data Utilization Status

This section examines browsing rate (bci/BC), download rate (dci/DC), and their comparative analysis. First, browsing volume (bci) and download volume (dci) for each region are statistically analyzed, as shown in Figure 7 [Figure 7: see original paper]. Meanwhile, charts of average browsing rate (bci/nci) and download rate (dci/nci) for single samples and average browsing rate ( ) and download rate ( ) for overall samples are drawn, as shown in Figure 8 [Figure 8: see original paper].

3.2.1 Browsing Rate of Open Data from Regional Perspective Similar to the thematic perspective, line charts and scatter plots of browsing rates for each region are drawn together, as shown in Figure 9 [Figure 9: see original paper], to facilitate observation of attention levels and status of regional open data. The X-axis represents the six cities, i.e., ci (i=1,2,…,6).

Combined with Figure 8(b) and Figure 9, whether by regional browsing rate or overall sample average browsing rate, Shanghai and Guiyang perform best. Taking Shanghai as an example, its data browsing rate is 166.883 times that of Jinan, while its overall sample average browsing rate reaches 0.63, far above the average (0.1667). Jinan shows the opposite, ranking last in both indicators, with other cities ranking slightly differently. Additionally, the sum of browsing rates for transportation theme data opened by Shanghai and Guiyang accounts for 89.3% of the total browsing rates for transportation data across all six cities.

Furthermore, combined with the browsing rate scatter plot, cities with high browsing rates like Shanghai and Guiyang also have larger spans in their scatter plots, with more data points falling above the average (1.67). Apart from Shanghai and Guiyang, other cities rank as follows: Wuhan, Guangzhou, Harbin, and Jinan. Although Harbin's overall browsing rate is not high, its corresponding scatter plot spans a relatively large range, and there are a considerable number of data points on its government data open platform with browsing rates above the average.

3.2.2 Download Rate of Open Data from Regional Perspective Analyzing the download rates of open data across regions helps analyze user utilization of each city's open data. Line charts and scatter plots of download rates for each region are shown in Figure 10 [Figure 10: see original paper].

As shown in Figures 8 and 10, Shanghai and Guiyang have high download rates as a whole, and their sample average download rates also hold significant advantages. For example, Shanghai, with the highest open data download rate, differs from the lowest, Jinan, by as much as 136 times. Additionally, Shanghai's nine theme datasets (excluding social security and employment) all have the highest download rates, all exceeding the average (approximately 0.167), with education and technology reaching as high as 0.798. However, some cities have low open data download rates, such as Wuhan, which ranks last in all download rate indicators.

Combined with the scatter plots of open data in each city, similar to browsing rates, cities with high download rates have larger spans in their scatter plots. Conversely, cities with low overall download rates have relatively smaller spans, but this does not mean that browsing rates and download rates change proportionally.

3.2.3 Utilization Status of Open Data from Regional Perspective To further reveal the utilization status of government open data from a regional perspective, hierarchical clustering is used to analyze the utilization status of open data in each region based on four indicators including browsing rate. Squared Euclidean distance is used for individual distances, and between-group average linkage is used for inter-cluster distances. The final clustering results are shown in the icicle plot and dendrogram in Figure 11 [Figure 11: see original paper].

Due to the small number of regional research samples, the observation method is directly used to determine the number of clusters. According to the icicle plot, when clustered into 4 categories, Harbin and Guangzhou form one category, Jinan and Wuhan form one category, and Guiyang and Shanghai each form separate categories. When clustered into 3 categories, Harbin, Guangzhou, Jinan, and Wuhan form one category, while Guiyang and Shanghai each form separate categories. For more detailed analysis of the distribution of open data utilization, the six research samples are clustered into 4 categories.

The first category includes Harbin and Guangzhou. The two cities have high similarity and cluster first with a coefficient of only 0.043. In terms of the four measurement indicators, Harbin and Guangzhou have similar rankings, following Shanghai and Guiyang, with most indicators in the middle range. The similarity between Harbin and Guangzhou's open data is also reflected in their data opening start time, quantity, and format. Although they started relatively late, the attention and utilization of their government data are acceptable.

The second category includes Jinan and Wuhan. The coefficient between these two cities is 0.125, indicating high similarity. Although Wuhan started opening data two years earlier than Jinan, with a large total number of open datasets and high browsing rate, its overall and average download rates are very low due to factors such as open format. Similarly, Jinan started late with fewer datasets, ranking at the bottom in all four indicators. Therefore, the gap with Wuhan is not obvious, and these two cities cluster together.

The third category is Guiyang. Although Guiyang started opening data relatively late, it achieved second place in the open data index within less than a year [24]. Except for the overall sample average browsing rate, all other indicators of Guiyang's open data exceed the average. Additionally, the browsing volume of its nine theme datasets (people's livelihood services theme is temporarily missing) consistently ranks in the top three. Guiyang's open data attention and utilization degree are second only to Shanghai.

The fourth category is Shanghai. Shanghai's open data browsing and download rates both rank first. Since opening data in 2012, Shanghai has maintained its position in the first tier of local government open data through government guidance, improved data quality, emphasis on user participation, and data innovation [25]. In this study, Shanghai's open data ranks first in all four indicators including browsing rate and download rate. The single average download rate of its economic and industrial commerce data is as high as 352 times per entry (approximate), more than double that of other categories.

4. Correlation Analysis Between Browsing Rate and Download Rate

4.1 Overall Correlation Analysis Between Browsing Rate and Download Rate of Open Data

Browsing rate and download rate represent user attention degree and data utilization degree respectively. Clarifying their relationship helps improve data utilization efficiency. First, without distinguishing research perspectives, R is used to plot a scatter diagram of browsing rate and download rate, as shown in Figure 12 [Figure 12: see original paper].

As shown in Figure 12, a large number of scatter points representing browsing rate and download rate are distributed on both sides of the trend line, with only some data showing the trend that higher browsing rates correspond to higher download rates. To further analyze the relationship between browsing rate and download rate, an R regression model is used to fit the data. First, linear least squares regression analysis is performed. The results show that the p-value of the regression coefficient (0.166, <2e-16) is very small, significantly $ 0; * * *alsoindicatesverysignificantsignificance.Meanwhile, theF - statistic = 3799, p-value :< 2.2e-16ismuchsmallerthan0.05, indicatingthattheoverallregressionmodelissignificantana of - fitR^{2}$ = 0.4058 < 0.5 indicates weak fitting, and regression diagnostic plots are drawn accordingly.

The four diagnostic plots in Figure 13 [Figure 13: see original paper] are: (1) residual vs. fitted plot, showing no obvious curved relationship; (2) residual Q-Q plot, indicating that experimental data do not follow a normal distribution; (3) standardized residuals vs. fitted plot, where the vertical coordinate is the square root of standardized residuals—larger residuals correspond to higher point positions, showing equal variance of model residuals; (4) residual vs. leverage plot, identifying outliers, high leverage points, and influential points.

In summary, although browsing rate and download rate are correlated overall, the correlation is weak.

4.2 Correlation Between Browsing Rate and Download Rate from Thematic Perspective

Although open data browsing rate and download rate are correlated overall, this does not indicate the degree of association for each theme. Therefore, regression analysis is performed separately for the 10 themes to obtain Table 3 , and scatter plots of browsing rate vs. download rate for each theme are drawn, as shown in Figure 14 [Figure 14: see original paper], to analyze their variation patterns under different themes.

Combined with Table 3 and Figure 14, the closer the R^2 value is to 1, the stronger the correlation between browsing rate and download rate, meaning higher browsing rates correspond to higher download rates, and also indicating a stronger association between data attention degree and utilization degree.

From the thematic perspective, eight major categories show varying degrees of positive correlation between browsing rate and download rate. First, education and technology and people's livelihood services themes show strong positive linear relationships, with R^2 coefficients indicating that the regression relationship can explain 86.33% and 83.05% of the variation in the dependent variable respectively, showing good regression effects. Among them, the browsing rate and download rate of education and technology data are basically proportional, with evenly distributed scatter points. Second, the regression correlation coefficient for economic and industrial commerce is 0.7044. Additionally, the R^2 values for social security and employment and energy and environment are both 0.01528, indicating that these two themes show no correlation between browsing rate and download rate, with opposite development trends. Most other themes show weak correlation.

Therefore, the degree of association between browsing rate and download rate from strongest to weakest from the thematic perspective is: education and technology > people's livelihood services > economic and industrial commerce > medical and health > government agencies and social organizations > transportation > culture, sports, and leisure > public security > social security and employment = energy and environment.

4.3 Correlation Between Browsing Rate and Download Rate from Regional Perspective

To statistically analyze the degree of association between browsing rate and download rate for open data in each region, regression analysis is performed separately for the six cities, with results shown in Table 4 . Meanwhile, scatter plots for the six cities are drawn, as shown in Figure 15 [Figure 15: see original paper].

As shown in Table 4, from the regional perspective, browsing rate and download rate of open data in the six cities are positively correlated, but the correlation shows certain differences. Specifically, Jinan shows a strong positive linear relationship between browsing rate and download rate, with an R^2 coefficient of 0.7803, which can also be verified in the regional scatter plot. Shanghai follows closely, with a regression correlation coefficient of 0.7534 between download rate and browsing rate. Additionally, the squared correlation coefficients for Wuhan and Guiyang are 0.3183 and 0.1537 respectively, indicating that the regression relationship can only explain 31.83% and 15.37% of the variation in the dependent variable, showing poor regression effects. Therefore, the association between browsing rate and download rate for Wuhan and Guiyang is the weakest. In summary, the degree of association between browsing rate and download rate from strongest to weakest from the regional perspective is: Jinan > Shanghai > Guangzhou > Harbin > Wuhan > Guiyang.

5. Conclusions and Discussion

(1) Through cluster analysis, the following conclusions are drawn. First, from the thematic perspective, users pay higher attention to fields closely related to daily life such as economy and livelihood. The utilization status of data from 10 different themes shows certain differences, which are divided into 4 categories based on similarity: culture, sports, and leisure; social security and employment; and education and technology form the first category; economic and industrial commerce forms the second category alone; transportation and medical and health form the third category; people's livelihood services, energy and environment, government agencies and social organizations, and public security cluster into the fourth category. Second, from the regional perspective, different cities show different utilization degrees of open data, presenting obvious high-low distinctions: Harbin and Guangzhou form the first category; Jinan and Wuhan form the second category; Guiyang and Shanghai each form separate third and fourth categories.

(2) Through regression analysis, the variation patterns between browsing rate and download rate of open data are identified to explore open data utilization.

First, overall, browsing rate and download rate of open data are weakly correlated. This indicates that high values for both browsing rate and download rate rarely occur simultaneously across the 10 major themes. However, since both browsing rate and download rate are important indicators for measuring user utilization of open data, a strong positive correlation with simultaneously high values represents the optimal state of data utilization. Therefore, governments should actively take relevant measures to improve browsing and download rates.

Second, from the thematic perspective, themes related to social livelihood such as education and technology, people's livelihood services, and economic and industrial commerce show high utilization, with positive correlation between browsing rate and download rate. These data are naturally more highly concerned due to their close connection with public daily life. In the era of "Internet Plus" e-government, governments should continue to use their advantages to understand, analyze, and maximize user satisfaction. However, themes such as social security and employment and energy and environment show no correlation between browsing rate and download rate, with opposite development trends, while most other themes show weak correlation. Social security and employment, energy and environment, public security, and other themes are also closely related to users' lives, but their utilization status is not ideal, which is related to factors such as the number of open datasets and user demand. Therefore, it is recommended to analyze the causes of this phenomenon in detail, strengthen publicity and guidance for open data, and targeted improve the utilization status of such data. Meanwhile, the overall characteristics of open data from the thematic perspective are inconsistent with some individual fea-

tures, so while improving overall utilization, attention should also be paid to the value-added of internal data.

Finally, from the regional perspective, Jinan and Shanghai show positive correlation between browsing rate and download rate, with Shanghai ranking first and Jinan last in open data utilization. Other cities such as Guiyang, although having high browsing and download rates, show weak linear relationships. The utilization of open data by Chinese urban governments shows characteristics of imbalance. Additionally, the overall characteristics of open data are generally consistent with some features. The imbalance in open data utilization is influenced by multiple factors including development process, public awareness, utilization environment, and socio-economic development level (influencing factors will be analyzed in another paper). Therefore, governments should have open thinking and awareness, accelerate the pace of data openness, improve social awareness, and narrow the utilization gap among Chinese cities while improving open data utilization rates.

6. Limitations and Future Work

This study starts from a micro perspective, selecting partial data resources from government open data platforms in Harbin, Jinan, Shanghai, Wuhan, Guangzhou, and Guiyang as research samples. Based on classification, it calculates and analyzes indicators such as browsing rate and download rate from thematic and regional perspectives, and further conducts cluster and correlation analysis. It finds that the utilization status of open data shows imbalance across different themes and cities, and finally proposes relevant recommendations. Although there is no unified standard for measuring and evaluating the utilization status of open data, and reflecting its utilization status solely through browsing rate and download rate has certain limitations, compared with pure qualitative analysis or macro analysis, this study takes open data itself as the entry point, enhances persuasiveness and credibility through real-time data capture and quantitative indicator calculation, and provides significant reference for promoting open data utilization overall.

However, this study also has some limitations. First, since comprehensive sample analysis was not conducted, it may affect the comprehensiveness and scientific nature of the analysis. Second, only the static utilization status of open data was analyzed, without time series analysis in a dynamic environment. Finally, the factors influencing the utilization status of open data were not deeply explored. These limitations will be addressed in future research.

References

[1] Opengovernmentdata[EB/OL].[2017-12-13].https://opengovernmentdata.org/.
[2] Theopendatabarometer[EB/OL].[2017-12-14].http://opendatabarometer.org/4thedition/report/.
[3] PARYCEKP, CHTLJ, GINNERM. Opengovernmentdataimple-mentation[J]. Informationpolicy, 2014, 19(1): 73-217-240.

[4] MARTINC. Barrierstotheopengovernmentdataagenda: takinga multi-levelperspective[J]. Policy&internet, 2015, 6(3): 399-418.
[5] ROSEIRAC. Exploringthebarriersinthecommercialuseofopen government-data[J]. Governmentinformationquarterly, 2016, 33(3): 535-551.
[6] ALEXOPOULOSC, LOUKISE, CHARALABIDISY. Amethodologyfordeter-miningthevaluegenerationmechanismandimprovementprioritiesofopengovernmentdatasystems[J]. Computerscience&informationsystems, 2016, 13(1): 237-258.
[7] ATTARDJ, ORLANDIF, SCERRIS, etal. Asystematicreviewofopengovern-mentdatainitiatives[J]. Governmentinformationquarterly, 2015, 32(4): 399-418.
[8] ZELETIFA, OJOAC, CURRYE. Exploringtheeconomicvalueofopengovern-mentdata[C]//Internationalconferenceontheoryandpracticeofelectronicgovernance. The9thinternationalconfer-enceoninformation systems (ICIS2013): reshapingso-cietythroughinformationsystemsdesign. Milano: ICIS, 2013.
[9] JETZKET, AVITALM, BJØRN-ANDERSENN. Generatingvaluefromopen-governmentdata[C]//Conferenceoninformationsystems(ICIS2013). NewYork: ACM, 2016: 211-214.
[10] GRAVESA, HENDLERJ. Astudyoftheuseofvisualizationsforopengovern-mentdata[C]//Internationalconferenceontheoryandpracticeofelectronicgovernance. The9thinternationalconfer-enceoninformation systems (ICIS2013): reshapingso-cietythroughinformationsystemsdesign. Milano: ICIS, 2013.
[11] ZhengLei, GuanWenwen. Researchonopengovernmentdataevaluationframe-work, indicatorsandmethods[J]. LibraryandInformationService, 2016, 60(18): 43-55.
[12] ChenShuixiang. Researchonvalueevaluationofgovernmentdataopenplatformbasedonuserutilization—taking19localgovernmentdataopenplatformsasexamples[J]. InformationScience, 2017, 35(10): 94-98, 102.
[13] WuLin, WuShiyu. Constructionandapplicationofserviceperformanceeval-uationsystemforurbanopengovernmentdataplatform[J]. LibraryTribune, 2018, 38(2): 59-65.
[14] LiuXinping, XiaoXin, HuangYiyi. Currentstatus, problemsandcounter-measuresofenvironmentaldataopeningbyChineselocalgovernments: analysisbase-donopen dataplatformsofsomeprovincesandcities[J]. E-Government, 2017(9): 30-40.
[15] CaoYujia. Survivalstatusofgovernmentopendata: surveyreportfrom19localgovernmentsinChina[J]. LibraryandInformationService, 2016, 60(14): 94-101.
[16] WangLei, DengLingyun. Preliminarystudyongovernmentdataopenguaran-teemechanismfrombigdataperspective[J]. InformationStudies:Theory&Application, 2017, 40(2): 77-79.
[17] MaHaiqun, PanPan. Analysisofcurrentstatusofdomesticandforeignopendat-apolicyresearchandjudgmentofChina'sresearchtrends[J]. JournalofLibraryScien-ceinChina, 2015, 41(5): 76-86.
[18] WangFashuo, WangXiang. Influencingfactorsandimplementationpathofgovernmentdataopeningandutilizati Aqualitativestudybasedongroundedtheory[J]. JournalofIntelligence, 2016, 35(7): 151-157.
[19] HuangRuhua, WangChunying. Investigationandanalysisofdatamanage-mentfunctionsofgovernmentdataopenplatformsinUKandUSA[J]. LibraryandIn-

formationService, 2016, 60(19): 24-30.
[20] ChenMei. Governmentdataopeningandutilization: connotation, progressandimplications[J]. LibraryDevelopment, 2017(9): 44-50, 77.
[21] XuHuina, ZhengLei. User-orientedopengovernmentdataplatform: comparativestudybetweenNewYorkandShanghai[J]. E-Government, 2015(7): 37-45.
[22] ZhangZiliang, MaHaiqun. Comparativestudyonutilizationeffectofgovernmentdataopenplatformsin China[J]. DigitalLibraryForum, 2016(6): 8-15.
[23] YaoJiayi. DataWarehouseandDataMiningTechnology: PrinciplesandApplications[M]. Beijing: ElectronicIndustryPress, 2009.
[24] "2017ChinaLocalGovernmentDataOpenPlatformReport"[EB/OL].[2017-05-28].http://www.cbdio.com/BigData/2017-05/28/content_{5528780}.htm.

Author Contributions

DuanYaoqing: Proposedresearchideas, revisedpaper;
QiuXueting: Datacollectionandanalysis, wrotepaper;
HeSiqi: Datacollectionandanalysis.

Note: Figure translations are in progress. See original paper for figures.