Text Mining-Based Comparative Study of Provincial Government Open Data Platforms in China (Postprint)

Authors: Chen Mei, He Qi

Date: 2023-04-01T15:51:25+00:00

Abstract

[Purpose/Significance] This study takes 14 provincial government open data platforms in China as research objects, conducting comparative analysis from multiple dimensions to provide reference suggestions for the development of government open data platforms in China.

[Method/Process] Data were collected through web scraping technology, subjected to descriptive analysis, and text mining was performed using the Tf-idf model. From the perspectives of data layer and platform layer dimensions, qualitative and quantitative analysis methods were employed to compare aspects including data resource granularity, domain distribution, timeliness, format types, search types, access conversion rates, and user feedback.

[Results/Conclusion] Currently, the development levels of open data platforms vary across provinces, with certain room for improvement. For example, the dataset publication scheme should be comprehensively considered based on provincial characteristics and dataset volume. During the construction process, attention should be devoted to aspects such as open platform data retrieval methods, training initiatives, and user feedback.

Full Text

A Comparative Study of Provincial Government Open Data Platforms in China Based on Text Mining

Chen Mei[1],[2], **He Qi**[1],[2]

[1]School of Public Administration, Zhongnan University of Economics and Law, Wuhan 430073
[2]National Governance and Public Policy Research Center, Zhongnan University of Economics and Law, Wuhan 430073

Abstract: [Purpose/Significance] This study examines 14 provincial government open data platforms in China, conducting a comparative analysis across multiple dimensions to provide recommendations for the development of government open data platforms in China. [Method/Process] Data were collected through web scraping techniques and subjected to descriptive analysis and text mining using the TF-IDF model. Starting from the dimensions of data layer and platform layer, qualitative and quantitative methods were employed to compare data resource granularity, domain distribution, timeliness, format types, retrieval types, access conversion rates, and user feedback. [Result/Conclusion] Currently, provincial open data platforms are at varying stages of development and have room for improvement. For instance, dataset release plans should be formulated based on provincial characteristics and dataset volume, with attention paid to data retrieval methods, training programs, and user feedback during platform construction.

Keywords: open data; government open data; open government; comparison

1. Introduction

Government open data platforms represent an effective means of enhancing government transparency. By opening government data and "showcasing" government data assets, these platforms not only safeguard citizens' rights to access government data publicly, thereby improving trust in government, but also increase user participation, interaction, and self-empowerment. As data with economic value for individuals and enterprises becomes available, social and economic growth can be stimulated. Although China has not yet established a unified national government open data platform, on February 9, 2021, the National Information Center issued the "Notice on User Questionnaire Survey for National Public Data Open Platform Construction," soliciting opinions and suggestions from all sectors of society regarding open platforms at all levels and specific needs for national public data openness. This initiative aims to further improve user experience, promote alignment between data supply and demand, and unlock greater benefits from data openness [**?**].

Existing research has primarily employed comparative and literature analysis methods, with less emphasis on text mining approaches. Previous studies can be categorized into three main types: (1) National-level government open data platforms. For example, Yang Ruixian et al. selected typical countries including the United States, United Kingdom, Japan, and Australia for comparison across three aspects: policy systems, safeguard mechanisms, and disclosure systems [**?**]. Wu Gang and Zeng Liying examined government open data platforms in the US, UK, Australia, Canada, as well as Beijing and Shanghai in China, exploring current development status from perspectives of resource availability, organization and retrieval, and service modes [**?**]. (2) Provincial and municipal government open data platforms. Tan Biyong and Chen Yan studied the quality of open

data platforms in 10 representative eastern, central, and western provinces and cities [**?**]. Yu Yihao and Li Weidong compared dataset quantities and API call frequencies across 10 provincial and municipal platforms from four functional perspectives—data, interface, application, and interaction—to analyze current conditions and problems, proposing optimization strategies [**?**]. (3) City-level government open data platforms. Deng Shengli and Xia Sudi compared resource quantities, visit volumes, and publication time distributions across eight Chinese and American cities from data layer and platform layer perspectives, highlighting the importance of focusing on people's livelihood data and optimizing user experience [**?**].

While existing research combines qualitative and quantitative methods to compare open data platforms across different countries and regions, few studies specifically focus on provincial government open data platforms using text mining methods. Although some research addresses provincial platforms, these typically mix provincial and municipal platforms in their comparisons. Given that administrative levels differ between national, provincial, and municipal platforms, and that data resource quantities and platform scales vary, the applicability of previous findings to provincial platforms requires further examination. Therefore, dedicated comparative studies of provincial government open data platforms are necessary to summarize successful experiences, identify potential issues, and provide references for the development of local government open data platforms and the construction of a national platform.

2. Comparative Analysis Framework

Drawing on the classification perspective of Deng Shengli and Xia Sudi [**?**], this study develops a comparative framework from two dimensions: data layer and platform layer. Data resource granularity, domain distribution, timeliness, and format types reflect the breadth and depth of government data openness and serve as indicators for the data layer. Retrieval types, access conversion rates, and user feedback reflect user-platform interaction and serve as indicators for the platform layer. The specific comparative framework is shown in Table 1 .

It should be noted that secondary indicators include datasets, APPs, and APIs. Datasets refer to collections of processed raw data, with their quantity directly reflecting platform development level. APP quantity indicates dataset usability, while API quantity reflects dataset openness and value extraction potential. Therefore, this study uses datasets, APPs, and APIs comprehensively to represent data layer resources.

3. Sample Selection and Research Methods

This study references the 18 provincial evaluation objects from Fudan University's Digital and Mobile Governance Laboratory's "China Local Government Data Open Report (Indicator System and Provincial Benchmarks)" released in October 2021. Data were collected from platforms between September 10-

15, 2021. After screening for website validity and data collection feasibility, 14 provinces (autonomous regions and municipalities) were selected as research objects: Hunan, Shandong, Shaanxi, Jiangxi, Ningxia Hui Autonomous Region (hereinafter "Ningxia"), Henan, Zhejiang, Hainan, Fujian, Guangdong, Guangxi Zhuang Autonomous Region (hereinafter "Guangxi"), Guizhou, Hebei, and Sichuan (in no particular order).

Several caveats apply to the data: (1) Platforms update data in real-time, so there may be minor lags and non-synchronous collection dates across provinces; (2) Some webpages were invalid or presented data in chart format, resulting in slight variations in the number of provinces included in different analysis modules; (3) Provinces with very few data records in certain modules were excluded to facilitate conclusion development without affecting results.

This study employs a mixed-methods approach combining qualitative and quantitative analysis. Python-based web scraping was used to collect raw data on datasets, APPs, and APIs. The TF-IDF model was applied to mine user feedback content from interactive columns, revealing more authentic and accurate insights (see Figure 1 [Figure 1: see original paper]). These analyses identify problems and propose reasonable recommendations to improve government open data platforms.

4. Comparative Analysis of Provincial Government Open Data Platforms

4.1 Data Resource Granularity

Data openness serves as the first stage for further research and knowledge innovation, directly impacting data production, dissemination, management, and usage. Proper data resource classification reduces users' time and labor costs, improving platform usability. This study measures granularity—the density of resource classification—using the formula: Data Resource Granularity = Resource Quantity / Domain Distribution, specifically through dataset granularity, APP granularity, and API granularity.

Given large variations in similar data resources, median granularity values serve as the comparison standard: values above the median indicate overly coarse classification, while values below indicate overly fine classification. Table 2 presents granularity statistics for the 14 provinces, with visualization shown in Figure 2 [Figure 2: see original paper].

Figure 2 reveals significant imbalances in granularity across provincial platforms, with most classifications being unreasonable. Provinces with overly coarse classification should reconsider their categories by subdividing broad domains, though this may be challenging for large datasets. Therefore, only new datasets could be reclassified. Platforms with overly fine classification may maintain current status during this developmental phase and reconsider classification once datasets stabilize.

4.2 Domain Distribution

Domain distribution refers to the proportion of datasets belonging to different domains. Combined with provincial economic levels, education, and policy directions, domain proportions reflect platform priorities and focal points during specific periods. The analysis identified 79 domains across 14 platforms, with issues including: (1) Similar domains (e.g., "cultural leisure" vs. "culture"); (2) Different terms for identical content (e.g., "safety supervision" vs. "safety monitoring"); (3) Ambiguous categories like "none" or "other."

To address these, similar domains were merged and "none/other" categories were consolidated into an "unassigned" domain. The 79 domains were reorganized into 30 categories including government affairs, resources/energy/environment, intellectual property, healthcare, culture/sports/leisure, statistical services, market regulation, food/drug safety, ecological environmental protection, industry/agriculture, etc. (see Table 3 and Figure 3 [Figure 3: see original paper]).

Literature indicates that open data platforms should adopt a user-oriented approach [**?**], though political orientation reflected in data cannot be ignored. Recent hot topics such as "carbon neutrality," "double reduction," "digital RMB," "ESG," and "COVID-19" reflect current government priorities. Figure 3 shows that open data primarily concentrates on people's livelihood, market regulation, economic/finance, education, safety, institutional groups, resources/energy/environment, healthcare, industry/agriculture, and urban/rural development—aligning closely with national strategic directions and balancing data usability with user needs.

4.3 Timeliness

In the digital era, timeliness—a core principle of open data—largely determines data quality and significantly impacts user satisfaction and government trust, especially during emergencies [**?**]. Conversely, "outdated" data represents "invalid" data with limited practical value and may cause database overload. Governments should therefore prioritize timely data updates while backing up and cleaning outdated data to reduce storage pressure and improve platform efficiency.

Table 4 shows significant variations in timeliness across platforms. Provinces with >5,000 datasets show consistent annual growth; those with 500-5,000 datasets show an initial rise followed by decline; those with <500 datasets maintain overall upward trends. Early growth reflects platform establishment from "0 to 1," while subsequent decline may indicate stabilization. Notably, despite Ningxia's late platform launch, it shows upward trends. Shaanxi's rapid 2018 development was followed by minimal dataset releases, possibly due to focus shift toward public services and credit domains where most data are annual. Thus, even with moderate dataset volumes, platforms may experience rise-then-decline patterns, typically peaking at year-end.

4.4 Format Types

Government open data platforms continuously expand their datasets, which are stored in various formats. Richer format types indicate higher openness. As shown in Figure 4 [Figure 4: see original paper] and Figure 5 [Figure 5: see original paper], platforms primarily use XLSX, JSON, XML, and CSV, with 5-8 format types being optimal. Hainan, with only 233 datasets, uses just XLS. Guangdong and Sichuan have similar dataset volumes but differ significantly in format variety. PDF, TXT, and DOC formats appear only sporadically and can be disregarded. When planning format types, similar formats (e.g., DOCX/DOC, XLSX/XLS) should be avoided to reduce system burden.

4.5 Retrieval Types

According to World Bank definitions, open data must meet two conditions: (1) Legally open, with explicit licenses permitting commercial/non-commercial use and unrestricted reuse; (2) Technically open, provided in machine-readable standard formats retrievable and processable by common applications. Retrieval type variety thus reflects openness level [**?**]. Unreasonable retrieval configuration reduces search efficiency, user satisfaction, and may even prevent data access, increasing development and maintenance costs.

Table 5 summarizes retrieval capabilities across platforms. Common methods include keyword search, directory browsing, department jurisdiction search, domain search, map services, format search, location search, time range search, and open method search. Provinces with >800 datasets typically offer 8+ retrieval types, while those with <800 datasets offer 4-5 types. Retrieval methods should be determined based on dataset volume, with combined approaches leveraging complementary advantages to maximize efficiency.

4.6 Access Conversion Rate

Access and download behaviors represent user-platform interaction. While access indicates interest, downloads reflect actual demand. The access conversion rate (downloads/visits) comprehensively measures dataset "popularity" and "attractiveness." This metric embodies not just a linear value indicator but a "spiral" feedback mechanism: governments can gauge user needs, validate demand assessments, and adjust strategies accordingly. Higher conversion rates indicate better alignment with user needs.

However, artificially high rates may occur when both downloads and visits are low, requiring outlier removal or post-hoc verification. For instance, Hainan's small dataset volume yields high visit and download volumes, suggesting government failure to adequately consider user needs, resulting in unclear release directions and overlooked data demands.

Visits and downloads correlate linearly with dataset quantity across domains—domains with more datasets receive more visits and downloads, and users typi-

cally download accessed datasets. While data quality generally meets user needs, some provinces require improvement. For example, Fujian shows poor consistency between visits and downloads, indicating that while users are interested in certain domain data, content quality fails to meet their needs, preventing downloads.

4.7 User Feedback

Government open data platforms increase opportunities for public participation. Interactive columns can overcome the "poor connection" between ordinary users and public affairs, capturing valuable user experience feedback. This study categorizes user feedback into four modules: "Data Application," "Error Correction," "Opinion Feedback," and "Consultation" (see Table 6).

Using the TF-IDF algorithm with Harbin Institute of Technology's stopword list and custom terms (punctuation, polite official responses, and province names), Jieba segmentation extracted top 20 terms. Word clouds were generated by category (title term frequency, problem description, reply term frequency) to identify themes and trace original texts for accurate interpretation.

TF-IDF assesses term importance: terms frequent in a text but rare in the corpus are most important. This filters irrelevant terms (e.g., "notice," "Qingdao," "so"), improving content authenticity. Though it cannot distinguish polysemy, its suitability for objective rather than emotional texts makes it appropriate here. Resulting keyword word clouds are shown in Figures 7 [Figure 7: see original paper] through 9 [Figure 9: see original paper].

For users, having their feedback adopted and resolved significantly boosts participation, creating a virtuous cycle that lays a solid foundation for platform development.

5. Conclusions and Recommendations

5.1 Data Layer

5.1.1 Data Domain Classification Current domain classification in Chinese provincial government open data platforms is poorly rationalized. Improvements should consider dataset quantity, current policies, and provincial characteristics to avoid classifications that are too coarse or too fine, which increase maintenance costs and search difficulties.

5.1.2 Data Format Selection Current dataset format configurations are generally reasonable. Future platforms should adopt XLSX, JSON, XML, and CSV as baseline formats while avoiding functionally similar formats. Additional formats should be selected based on needs, collection difficulty, and data characteristics.

5.1.3 Data Content Release Current platform content aligns well with policy directions but 仍有改进空间. First, dataset release should consider access conversion rates, prioritizing high-demand domains and potentially excluding low-interest datasets. Since user interests are dynamic, governments should regularly track and catalog datasets, emphasizing "targeted" releases to improve practical value. Second, data "freshness" must be maintained through timely updates and proper backup of old datasets, strengthening data quality governance and improving user satisfaction.

5.2 Platform Layer

5.2.1 Retrieval Method Usage Information retrieval should be comprehensive, flexible, and efficient. Provinces should configure retrieval types based on dataset volume and combine multiple methods to improve search efficiency while reducing maintenance costs.

5.2.2 Training Program Implementation Governments or non-profit organizations should provide platform usage training through online channels like TikTok and Weibo, and compile user guides. For example, unlike North America and Europe, African countries have limited access to open government data and correspondingly limited capacity to leverage technology for growth. Code for Africa (CFA), the largest civic technology and data journalism lab network in Africa with teams in 20 countries/regions, aims to develop community technical skills and coding capabilities, creating opportunities for citizens to monitor government, enterprises, and public institutions. This organization not only views open data as a potential public asset but has also developed a data fellowship program embedding data-skilled personnel in various media and non-profit organizations.

5.2.3 User Feedback Improvement When responding to user feedback, staff should provide templated, specific answers aimed at genuinely solving problems. Vague responses create poor user experiences and negative perceptions of feedback as a waste of time. Based on word cloud themes, future improvements should focus on:

(1) **Precise and Timely Content**: Users currently demand data on education/culture and industry resources, such as college entrance exam scores, reservoir resources, and tourism data. However, dynamic user needs require governments to integrate feedback, current events, and national policies for timely adjustments, ensuring data drives economic and social development.

(2) **Simple and Convenient Functions**: Technical issues like real-name authentication failures exist. Some reported problems are actually normal operations, with conflicts arising from platform instability, browser incompatibility, or user error. Recommendations include: improving data governance for authenticity and accuracy; strengthening user guidance

through multi-channel promotion, function explanations for new pages, interface optimization, and clear query paths; and adopting a user-centered approach with specialized pages for vulnerable groups (e.g., elderly mode, disability mode) to enhance experience.

(3) **Professional and Specific Responses**: Vague replies like "system anomalies require technical maintenance" are common but counterproductive. Specific response templates should include problem description, solution method, resolution time, handling institution, and complaint email. Crucially, governments should conduct regular follow-up checks on resolved issues.

References

[?] Notice on User Questionnaire Survey for National Public Data Open Platform Construction [EB/OL]. [2021-10-25]. http://www.sic.gov.cn/News/612/10773.htm.
[?] Yang Ruixian, Mao Chunlei, Zuo Ze. A Comparative Study of Government Data Openness at Home and Abroad [J]. Journal of Intelligence, 2016, 35(5): 167-172.
[?] Wu Gang, Zeng Liying. A Comparative Study of Open Data Platform Construction at Home and Abroad [J]. Information and Documentation Services, 2016(6): 75-79.
[?] Tan Biyong, Chen Yan. Research on Data Quality of China's Open Government Data Platforms—A Study of Ten Provinces and Cities [J]. Journal of Intelligence, 2017(11): 99-105.
[?] Yu Yihao, Li Weidong. Current Status, Problems and Optimization Strategies of Local Government Data Open Platforms in China—A Study Based on 10 Local Government Data Open Platforms [J]. E-Government, 2018(10): 99-114.
[?] Deng Shengli, Xia Sudi. A Comparative Study of Chinese and American Urban Government Open Data Platforms [J]. Library Journal, 2019, 38(6): 57-68, 75.
[?] Wu Qunying, Ma Lei. Investigation on the Current Status of Provincial Government Open Data Platform Construction in China [J]. Information Research, 2020(9): 69-75.
[?] Chen Shuixiang. Value Evaluation Research on Government Data Open Platforms Based on User Utilization—A Case Study of 19 Local Government Data Open Platforms [J]. Information Science, 2017(10): 94-98, 102.
[?] Wang Qingyi, Gao Jie. Research on User-Oriented Services of U.S. Government Open Data and Its Implications—A Case Study of the U.S. Data.gov Website [J]. Journal of Intelligence, 2016(7): 145-150.
[?] Zhang Linxuan, Chu Jiewang, Cai Xiang, et al. Analysis on the Development Status and Countermeasures of Prefecture-Level Government Data Openness in China—A Case Study of Anhui Province [J]. Technology Intelligence Engineering, 2021, 7(4): 79-92.

Author Contributions:
Chen Mei: Research design, paper writing;
He Qi: Data collection, paper writing.

Acknowledgments: This work was supported by the National Natural Science Foundation of China project "Research on User-Oriented Open Government Data Usage Behavior Mechanisms and Privacy Risk Control" (Project No. 72004056) and the Central University Basic Research Fund of Zhongnan University of Economics and Law "Open Government Data Policy Optimization Research" (Project No. 2722022BQ039).

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