A Multi-source Data Fusion Analysis Model in the Research on Public Sentiment Index

Yunya Zhu1,*, Zhengang Zhai1, Bingtao Gao1, Song Xu1 and Nannan Wu1
1The 36th Research Institute of CETC, Jiaxing, China

*Corresponding author e-mail: zhuyy@jec.com.cn

Abstract. At present, China’s economy is continuously developing and making progress. The deep integration of society and big data artificial intelligence and other technologies is constantly promoting the construction of new smart cities. The major livelihood service systems have provided great convenience to the people. At the same time, they have accumulated massive livelihood data. Mining the value of data to solve people’s main problems has been widely concerned by the government and the community. However, the data structure of each system is different. The data is noisy and the systems cannot be interconnected, which leads to the one-sided and single results of data analysis, and it is impossible to use these data to obtain effective results and conclusions. This paper proposes a Public Sentiment Index that reflects the intensity of public appeals. A new multi-source data fusion analysis model is proposed by using a multi-source data fusion method. By constructing a public sentiment database and a public sentiment index model, it can truly reflect the changes in people's demands. This research can provide a reliable and scientific basis for the government to improve decision-making capacity and service quality.

1. Introduction

The government and all walks of life always focus on the issue of public sentiment. More and more scholars are focusing on comprehensive social assessments [1,2,3]. At present, some scholars have introduced the European social development quality theory and applied it to evaluate the development of different cities in China [4,5,6,7]. These studies have paid too much attention to the objectivity of the indicators which are difficult to obtain, and lacked in considering the complexity of the actual data. It is difficult to apply to the construction of smart cities directly. The "Action plan for big data development" issued by the State Council in 2015 clearly proposed that big data technology integrated into urban governance provided a new way to improve urban governance capabilities [8].

This paper adopts multi-source data fusion technology to fuse data from different data sources with different structures to effective classification and sorting, further construct a public sentiment index model, and quickly sense the intensity of public demand. The experiments on the data of F-city show that the method proposed in this paper can promote the solution of popular demands and has practical significance for advancing the construction of smart cities.
2. Multi-source data fusion analysis technology

2.1. The connotation of multi-source data fusion

Multi-source data fusion refers to the use of different algorithmic tools and data technologies to effectively collect, sort, investigate, and analyse relevant data types, structures and values, so that data from multiple sources are fused together, and multiple types of data are merged. It is a process of obtaining high-value information resources through scientific and objective evaluation and analysis of various types of data. The purpose of multi-source data fusion is to fully optimize the processing of data of different types, structures and contents, to take advantage of data from different sources, and to extract a unified structure feature from massive data, which is more reliable and valuable than a single data. This method facilitates decision management and makes the data of service users more scientific.

2.2. Multi-source data fusion technology and multi-source data fusion system

From the current research, multi-source data fusion technology is a data processing technology that can comprehensively analyze, calculate, mine and manage data from multiple sources. Through comprehensive mining, analysis, and integration of data resources, the data becomes an interconnected organic whole, which is convenient for information analysts to fully grasp the data. Multi-source data fusion can be divided into three levels: data-level fusion, feature-level fusion, and decision-level fusion. In this paper, data-level fusion technology is adopted to build a public sentiment database and clean data from public sentiment system platforms such as Mayor’s Hotline, Safe City, and integrated law enforcement.

2.3. Multi-source data fusion model

A data-level fusion method is used in this paper, which described as Equation 1.

$$DR_{(i,j)} = \sum_{i=1}^{N} W_{(i,j)} D_{(i,j)}$$

Among them, $DR_{(i,j)}$ indicates the value of the data fusion calculation result. $i \in (1, 2, \cdots, N)$ indicates the classification of the problem to be solved. $j \in (1, 2, \cdots, M)$ indicates the number of data sources. $W_{(i,j)}$ indicates the weight of the j-th data source for the problem i. $D_{(i,j)}$ indicates the value of the j-th data source for the problem i.

2.4. Practical significance of multi-source data fusion

Multi-source data fusion technology can satisfy the needs of people to obtain data from multiple channels and diversify their data usage. It is a practical result of the development of information technology. Data fusion technology is fundamentally different from traditional data usage methods. With the help of data distributed processing systems and data identification systems, multiple data sources can be accurately identified, and data fusion can be carried out according to the data structure, type analysis, and evaluation of different data sources. In social practice, multi-source data fusion technology can help people mine the value of data, discover the laws of data change, and grasp the internal correlation of data.

3. Building a public sentiment database

Public sentiment issues have always attracted the attention of all sectors of society. With the continuous development of new smart cities, relevant departments have established a series of
platforms which receive public sentiment information, such as the 12345 mayor hotline platform, the safe city system platform, and a comprehensive law enforcement platform. After a long run, various platforms have accumulated a large amount of raw data. However, these service platforms are not interconnected, so data cannot be shared. It is difficult to fully mining the value of data. This paper uses data-level fusion technology to fuse data from multiple systems into a unified structure.

3.1. Data cleaning

During the operation of each service platform, some information is collected which has nothing to do with people's needs. This information is useless for public sentiment analysis and may even interfere with the results of public sentiment analysis. In order to reduce data noise, this paper uses word segmentation technology of natural language processing to remove useless work record data which has nothing to do with public sentiment in each data source, obtains valid public sentiment data, and establishes public sentiment database.

First, Construct a list K which is based on the statistical results of experience and word frequency: if a word or phrase represents "effective public sentiment data," then the word or phrase is put into the list. Second, using the word segmentation technology of natural language processing, the content of the "problem description" is segmented by the Chinese word segmentation method to obtain a list T which is composed of multiple phrases. For each phrase item in the list, compare it with each item in the list K. If none of the items in the list K are a subset of the items, the "problem description" belongs to "useless public sentiment data". The specific process is shown in Figure 2.

![Figure 2. Data cleaning flow chart.](image)

3.2. Integrate the data of public sentiment

After cleaning the data, we obtained valid public sentiment data from those main systems. The common fields are extracted from the valid public sentiment data to form a public sentiment dataset.
According to each index classification criterion and public sentiment database, keyword matching is performed through Bayesian algorithm, and the similarity of paragraphs is matched using Jaccard algorithm. Then a classification model is constructed to classify the public sentiment dataset. Add "source" to each source. The surface of this field comes from the public opinion system, and statistics are divided.

4. Construction of public sentiment index

4.1. Definition of public sentiment index

The 12345 mayor hotline platform, the safe city system platform, and other platforms all have the characteristics of large data volume and low value density. Combined with the characteristics of data sources collected from different related platform systems, a public sentiment index is constructed to fully reflect the intensity of public demand. This article believes that the public sentiment index is a relative number that can comprehensively reflect the intensity of people's demands. According to data source information, people's demands can be divided into five categories: complaints and reports, consultations, issues, asking for help, suggestions and thanks. A detailed description of each category is shown in Table 1 below:

| Numble | Type name                  | Content                                                                 |
|--------|----------------------------|-------------------------------------------------------------------------|
| 1      | Reports and complaints     | Report violations of laws, regulations and disciplines by citizens, legal persons, state agencies and other organizations, and members; complaints about insufficient and inefficient services of government departments and their employees. |
| 2      | Consultations              | Consulting services such as government information, policies and regulations, work guidelines and other related information. |
| 3      | Issues                     | Reflect on issues related to people's livelihood and other issues related to urban governance. |
| 4      | Asking for help            | Help messages for individuals, families or groups when they encounter difficulties. |
| 5      | Suggestions and thanks     | Give thanks or provide advice for government services, and reflect public opinion. |

4.2. Construction of the public sentiment index system

For the five types of public demands, this article focuses on the issue of public response. According to the results of data cleaning and data fusion, the problems of public response are mainly concentrated in four aspects: residents' life, public services, ecological environment, political science and law. In order to build a more accurate indicator system, this article subdivides each aspect and builds a public sentiment indicator system that includes 4 primary indicators and 16 secondary indicators. The detailed content of the indicators is shown in Table 2.
Table 2. Detailed description of the public sentiment index system

| No. | First Grade | Second Grade | Index content |
|-----|-------------|--------------|---------------|
| 1   | residents' life | Employment | Employment information and other consultations |
|     |             | Housing     | Issues such as provident fund deposits, withdrawals, loans and other business consultations, provident fund related policy consultations, bundled sales of house parking spaces, “dual contracts”, comprehensive renovation of small towns, etc. |
|     |             | social security | Social security policy consultation, daily social security business operations, labor security rights protection, etc. |
|     |             | consumption | Issues such as Food safety, quality of houses and cars, bankruptcy of recharged consumer stores, refund of housing fund, etc. |
|     |             | education | Issues such as unfair educational resources, etc. |
|     |             | medical | Difficulty in seeking medical treatment, expensive medical treatment, tension between doctors and patients, and medical treatment |
|     |             | traffic | Passenger centres, railway stations, etc., unlicensed driving, driving in violation of regulations and carrying people in reverse, traffic congestion, etc. |
|     |             | infrastructure | Expressway changeover construction, sewage pipe network construction, public parking lot construction, railway station and other transportation hub reconstruction, etc. |
|     |             | culture and sports | Library and gym event information consultation |
| 2   | public service | city appearance and environment | Urban littering, diffuse odour, oil fume pollution, domestic garbage classification, etc. |
|     |             | air pollution | Reports of secret emissions of industrial waste gas |
|     |             | water pollution | Leakage of industrial sewage, random discharge of industrial sewage |
|     |             | noise pollution | Corporate noise, construction noise at night |
|     |             | legal aid | Legal advice on civil disputes |
| 3   | ecological environment | administrative reconsideration | Citizens are dissatisfied with the law enforcement and punishment of administrative law enforcement departments such as traffic police and urban management. They call for complaints to request coordination and maintain their rights through administrative reconsideration. |
| 4   | political science and law | Complaint reporting | Report citizens, legal persons, state organs and other organizations and members' violations of laws, regulations, and disciplines. Complaints about inadequate services and inefficiency in government departments and their staff. |

4.3. Index weight determination

This article uses the Delphi method to determine the five index weights of complaints and reports, consultations, issues, Asking for help, Suggestions and thanks. It is assumed that of each data source is same important here. We designed an expert scoring table to compare the emotional intensity of each type. We also organized meetings for experts scoring. The experts give extreme, strong, obvious, slightly or equivalent evaluations. Extreme, strong, obvious, slightly, and equivalent evaluations correspond to 5 points, 4 points, 3 points, 2 points, and 1 point respectively. Finally, we calculated the weights of the indicators.

Table 3. The results of using Delphi method to determine the weight of each index.

| No. | Index                  | index weight |
|-----|------------------------|--------------|
| 1   | Reports and complaints | 0.32         |
| 2   | Consultations          | 0.22         |
| 3   | Issues                 | 0.21         |
| 4   | Asking for help        | 0.20         |
| 5   | Suggestions and thanks | 0.05         |
4.4. Index model

Complaints and reports, consultations, issues, asking for help, Suggestions and thanks, five different types of events are weighted. Then we get public sentiment index, which reflects changes in the intensity of people's demands over time. (We assume that each data source is equal important.).

According to the actual situation, we set the reference period and calculate the public sentiment index by week.

\[ K_i = \frac{\sum q(i,j) p(j)}{\sum q(0,j) p(j)} \times 1000 \]  

(2)

5. An application on public sentiment index from F-city

In recent years, F-City has been committed to promoting the construction of a new type of smart city, and has established an integrated operation management center and a city big data center. The city’s big data center in F-city collects multi-source data resources such as government affairs, businesses and institutions, and the Internet information. It breaks down information barriers, speeds up data collection, cleansing and exchange, and maximizes the value of data. In order to understand the changes in public sentiment, this article fuses the multi-source data of the 12345 mayor hotline, a safe city construction system, a city management system from big data center.

We cleaned the data source from the four major public sentiment related platforms of F-city. The public sentiment data set of F-city was established. According to each index, the data of the data set is classified and the classification result is statistically selected. The table selects the data classification results from August 14, 2018 to November 13, 2018.

**Table 4:** Statistics of the classification results of each data source by type (unit: bar)

| Time         | Reports and complaints | Consultations | Issue       | Asking for help | Suggestions and thanks | Total   |
|--------------|------------------------|---------------|-------------|-----------------|------------------------|---------|
| 2018/08/14   | 109                    | 3235          | 5033        | 2706            | 1048                   | 12131   |
| 2018/08/21   | 119                    | 3845          | 8802        | 4393            | 1988                   | 19147   |
| 2018/08/28   | 108                    | 5983          | 12508       | 5785            | 4283                   | 28667   |
| 2018/09/04   | 176                    | 8628          | 14156       | 6638            | 5244                   | 34842   |
| 2018/09/11   | 144                    | 7776          | 13409       | 6793            | 5365                   | 33487   |
| 2018/09/18   | 124                    | 5232          | 9675        | 5166            | 3349                   | 23546   |
| 2018/09/25   | 111                    | 3856          | 6785        | 3511            | 1845                   | 16108   |
| 2018/10/02   | 126                    | 3758          | 6486        | 2783            | 1596                   | 14749   |
| 2018/10/09   | 72                     | 1644          | 2443        | 1544            | 609                    | 6312    |
| 2018/10/16   | 122                    | 3675          | 5075        | 2783            | 1262                   | 12917   |
| 2018/10/23   | 120                    | 3795          | 6235        | 3056            | 1254                   | 14460   |
| 2018/10/30   | 117                    | 4119          | 6673        | 3173            | 1273                   | 15355   |
| 2018/11/06   | 163                    | 3146          | 5364        | 2929            | 1347                   | 12949   |
| 2018/11/13   | 111                    | 4158          | 6499        | 3670            | 1086                   | 15524   |

The weekly public sentiment index value is calculated according to the public sentiment index calculation formula with August 14, 2018 as the base period. The change of public sentiment index is shown in Figure 3:
Table 5: Calculation of public sentiment index values by week

| No. | Time       | Weighted value | Index value (K) |
|-----|------------|----------------|-----------------|
| 1   | 2018/08/14 | 2397.11        | 1000            |
| 2   | 2018/08/21 | 3710.4         | 1547.863886     |
| 3   | 2018/08/28 | 5348.65        | 2231.291013     |
| 4   | 2018/09/04 | 6517.04        | 2718.70711      |
| 5   | 2018/09/11 | 6199.54        | 2586.25595      |
| 6   | 2018/09/18 | 4423.12        | 1845.188581     |
| 7   | 2018/09/25 | 3103.14        | 1294.533834     |
| 8   | 2018/10/02 | 2865.54        | 1195.414478     |
| 9   | 2018/10/09 | 1237           | 516.0380625     |
| 10  | 2018/10/16 | 2532.99        | 1056.684925     |
| 11  | 2018/10/23 | 2856.55        | 1191.664129     |
| 12  | 2018/10/30 | 3043.2         | 1269.528724     |
| 13  | 2018/11/06 | 2523.87        | 1052.880343     |
| 14  | 2018/11/13 | 3103.37        | 1294.629783     |

Figure 3: Changes in public sentiment index from August to November 2018

The data in Figure 3 shows that from August to September in 2018, the public sentiment index rose sharply. At this stage, F-City had an activity named large visit. This activity went deep among the masses, took the initiative to understand the demands of the people. Through the activity, public’s problems are vigorously resolved. People are more enthusiastic about reflecting problems. With the end of the large visit activity, the cumulative disadvantages of the old governance methods are gradually surfaced. Public demands have not yet been fully satisfied. The public sentiment index has gradually declined within half a month. Subsequently, F-City found the abnormal situation. They changed their governance method, and combined with the experience of activity. Then public sentiment index has increased steadily.

6. Conclusion

This paper proposes a new multi-source data fusion analysis model to obtain effective public sentiment data. This paper constructed a public sentiment database. We make public's demands into five categories: complaints and reports, consultation, issues, asking for help, suggestions and thanks. This paper constructed a public sentiment index system consisting of 4 primary indicators and 16 secondary indicators and determined the weight of each category by using the Delphi method. We construct a public sentiment index model, and calculate the public sentiment index of F-city. The
fluctuation of the curve is consistent with the time of major events in F-city, which can reflect the objecti
tive law of public sentiment. Aiming at the change of public sentiment index, combined with major
events in the process of urban governance, the crux of public sentiment change can be effectively ana-
yzed, and symptomatic measures can be taken to improve the government's governance
capacity, service quality, and promote the construction of Smart City.

7. Acknowledgments
In this paper, the research was sponsored by the Major R & D Plan of Science and Technology of Zhejiang (Project No. 2019C03099).

References
[1] Gu Yingchun, Theoretical Thinking on Social Index System, Sociological Study, 1997 (06): 110-114.
[2] Song Linfei, Index System and Evaluation of China's Well-off Society, Social Sciences in Nanjing, 2010 (01): 12-20.
[3] Ren Fubing, Wu Qingfang, Guo Qiang, Construction of Evaluation Index System for Low-Carbon Society, Jianghuai Tribune, 2010 (01): 124-129.
[4] Lin Ka, Social Quality Theory: A New Perspective to Study the Construction of a Harmonious Society, Journal of Renmin University of China, 2010 (02): 111-117.
[5] Zhang Haidong, From Development Road to Social Quality: Paradigm Transformation of Social Development Research, Xinhua Digest, 2010 (14): 16-18.
[6] Lin Ka, Liu Xiaoping, Mao Hui, Social Trust and Social Quality: Data Analysis and Evaluation of Zhejiang Social Quality Survey, The Journal of Jiangsu Administration Institute, 2010 (04): 63-69.
[7] Yuan Hao, Ma Dan. Subjective well-being from the perspective of social quality: based on empirical research in Shanghai, Jilin University Journal Social Sciences Edition, 2011 (04): 140-147.
[8] Meng Tianguang, Zhao Juan, Big Data-Driven Intelligent Social Governance: Theoretical Construction and Governance System, Electronic Government, 2018 (08): 2-11.
[9] Ren Dong, Xia Yifan, Sun Yi, Research on People ’s Livelihood Satisfaction Index——Taking Sichuan ’s People ’s Livelihood Satisfaction Index as an example, Finance Science, 2015: 111-120.
[10] Zhou Zhaojie, Wang Hongchuan, Su Yang, How Can Chinese People Have a Higher Level of Happiness——Based on the Survey of China's Livelihood Index, Management World, 2015: 8-21.
[11] Si Yi, Jia Wenmei, Sensitivity analysis of urban public safety risk assessment indicators, China Work Safety Science and Technology, 2014, 10 (11): 71-76.
[12] Yingchun Zhang, Research on the Construction of the People's Livelihood Index System in the Pastoral Area of Qinghai from the Perspective of Precision Poverty Alleviation, Journal of Southwest University for Nationalities (Humanities and Social Sciences Edition), 2017, 38 (01): 17-21.
[13] Noll, H.H. Toward a European System of Social Indicators: theoretical framework and system architecture, Social Indicators Research, 2002, 58, 47-87.
[14] Kimitaka Tsutsumi, Kazutaka Shimada, Tsutomu Endo, Movie review classification based on a multiple classifier, Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation(PACLIC21),2008
[15] Zhang Weichong, Wang Fang, Zhao Hong, Multi-source information fusion for the identification of emerging technology development trends——Taking block chain as an example , Journal of The China Society for Scientific and Technical Information, 2019,38 (11): 1166-1176.