Cluster Comput
https://doi.org/10.1007/s10586-017-1348-8

Data mining and visualization of data-driven news in the era of big data

Erna Qi[1] · Xingrui Yang[1] · Zongjun Wang[1]

Received: 15 August 2017 / Revised: 31 October 2017 / Accepted: 6 November 2017

Abstract With the continuous deepening of study of data mining, the application area of data mining gradually expanded, its influence also spread to the media industry. Data visualization technology has changed the traditional narrative mode, make the news becomes a product that be produced. This paper analyzes the history of computer aided reporting to data news, the main models of data news visualization, and the process of data news production through data mining. The study found that data news focuses on the way of data processing in the entire news workflow, it involves not only classical computer graphics technology, image-processing technology and computer audio technology, but also more data analysis and visual processing technologies based on new media and cloud computing involved in. Research data mining and visualization of data-driven journalism can help journalists use big data to do news work better, deepen people's cognition of news events, find the logic which cannot be reflected in traditional news, and maximize the connotation of news report.

Keywords Data mining · Dig data · Data-driven journalism · News visualization analysis

1 Introduction

With the rapid development of computer technology and the wide application of data acquisition equipment, the ability to generate, collect, store and process data in various fields of society been greatly improved; the explosive growth of all kinds of data resources is one of the prominent characteristics. Not only a huge amount of information, but also available from a variety of sources, including by government agencies, social organizations, enterprises and public data acquisition, user data generated in the various media platform, the mobile terminal of the geographic information, the Internet of things in various state and so on. All the information that is generated from time to time becomes a torrent of data [1]. The New York Times also pointed out in a column: "the era of big data has come, in business, economics and other fields, decision-making will increasingly be based on data and analysis, rather than based on experience and intuition" [2]. The continuous accumulation of data is no more than a huge treasure, and when it accumulates to a certain extent, it will inevitably reflect some rules. However, the huge data is far beyond the ability of people's analysis and processing, and the knowledge contained in data resources not been fully exploited and utilized [3]. At the same time, people's demand for information is getting higher and higher, hoping to make a higher level of analysis and use. This promotes the development of data mining technology [4].

2 The outline of data mining and data mining application

2.1 Data mining research

Data mining, also known as knowledge discovery, appeared in the late 1980s. It is one of the most advanced research fields in the field of database and information decision making [5]. Researchers such as Fayyad believe that data mining is to obtain knowledge of people's interest from the database, which is implicit and latent. Experts and programmers import traditional decision support systems, knowledge and rules in

✉ Zongjun Wang
qierna14362qi@163.com

[1] School of Management, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China

Published online: 24 November 2017

the knowledge base that the decision maker can use the tools such as online analysis and processing directly. Data mining is to obtain information, relationships, trends and other information that not yet been discovered from a large number of internal databases [6]. Witten claimed that data mining also called data mining data mining is in accordance with the established business objectives from the mass of data to extract the advanced treatment of potential, effective and understandable model of process [7]. Chen said data mining at a very low level; it means use the existing database management system, query retrieval and reporting capabilities, combined with multidimensional analysis, statistical analysis, and then on-line analytical processing applied to obtain the statistical analysis for decision-making. In the deep level, data mining means the discovery of unprecedented, implicit knowledge from the database. They are all methods of extracting useful information from a database [8]. As far as decision support is concerned, the two are complementary. OLAP can be regarded as a generalized data mining method, whose purpose is to simplify and support online analysis, and the purpose of data mining is to automate the process as much as possible [9].

2.2 Data mining application research

Big data not only means a huge amount of information or data, but also means that the data processing, analysis, sharing, mining and other capabilities will be unprecedented upgrade; the exchange and mutual utilization of data among different industries and different fields have become very frequent. Data mining application research refers to the development of a variety of data mining systems and tools, and applications in various industries. In September 2008, nature launched a cover column called big data, telling the more important role that data plays in math, physics, biology, engineering, and social economy [10]. Typical application areas include: market analysis and forecasting; such as the ratings survey conducted by BBC broadcasting company, sales analysis and forecast of large supermarkets, sales channels and price analysis, etc. The structure of finance; statistical regression neural network prediction model, such as automatic investment system, can predict the optimal timing of investment; science research; Quake finder was used in the analysis of crustal tectonic activity; Web data mining; Web access pattern analysis, automatic classification, web content clustering; the engineering diagnosis. As a new method of knowledge discovery, data mining has also attracted the attention of engineering diagnosis. Many countries and research institutes have joined the research of data mining in the project of monitoring and diagnosis [11]. With the deepening of data mining research, the scale of data mining application field is gradually expanding, and the more notable are banking, entertainment/music, science and health care/HR in turn [12].

3 The development of data-driven news

At present, the focus and application of "big data" not only focus on the IT industry, marketing, public health and other fields, but also spread to the media industry [13]. With the advent of the big data era, the production mode of news is constantly innovating, and the data visualization technology is like an industrial revolution, which changes the traditional narrative mode, let news become products that be produced [14]. Mining depth information in fragmented unstructured data has become a key link for news media to provide information services, and thus spawned data news, or data driven news [15].

3.1 The development from computer-aided reporting to data driven news

Anderson points out that the basic idea of data news is not new, but could traced back to history [16]. From the evolution of news reporting form, data news is not a new form, and it has a continuous development relationship with precision news and computer aided news report [17]. In 1960s, American scholar and journalist Philip Meyer put forward the theory of precise journalism. Accurate news has the nature of in-depth reporting, especially in digital information. Computer aided reporting originates from the need for accurate news reporting. Data news can regarded as the further development and promotion of computer aided reporting in the era of big data. Compared with the accurate news and computer-aided reports, data news reports in the system, timeliness, interactivity and reading experience have made great progress. In addition, it also embodies the spirit of openness and sharing in the digital age. Massive data is freely accessible to users through the Internet, with advanced user centric tools, self-publishing, and crowdsourcing tools, it makes it easier for more people to use these data to make news [18].

3.2 Research on the concept of data news

Data news is a new achievement that the news media industry is constantly exploring and developing with the trend of data information. It is a new way of news production in the era of big data. In brief, it's data-driven reporting, specifically, by mining and displaying the relevance and model behind the data, using the rich and interactive visual communication, create a new way of news reporting. In the process of the development of data news, many scholars have defined it, Holland NU.nl. news website data researcher Jerry Vermanen believes that “Data news is a new set of skills that can be used to search, understand and visualize digital sources, data news can help news organizations achieve two important purposes: finding unique stories (not from news agencies) and performing watchdog functions”; Aron Pilhofer, a senior journalist at the New York Times, believes that data news is a new way of news narrative. Data journalism is a general term, it includes a set of still used in machine tools, techniques and methods of news narrative is increasing, ranging from the traditional computer aided report (using data as a “source”) to the forefront of the data visualization and application of a narrative way of news. Fang analyzed the data news from the perspective of presentation form, production process and Industry Development, defined as “data news” is a new type of news reporting method of data grabbing, mining, statistics, analysis and visualization” [19].

4 Research on visualization of data news

4.1 Visualization of data mining

Data visualization is a method of exploring, displaying and expressing data meanings by means of visual communication. It delivers all kinds of data and information visually, and data and information are the basis of design. It is a tool, with the progress of technology, the form of data visualization is more rich and diverse [20]. From the perspective of communication, the accessibility of visual language helps to eliminate the barriers between information communication and communication. In the context of big data, data news based on this principle, so that the original boring news and massive data hooked together. Under the simultaneous action of data analysis technology and visualization technology, the linking of data and news not only ensures the authenticity and accuracy of the data information, but also transmits the data information to the audience in a specific visual form [21] María Teresa described the process chart for visualizing data mining for us [22]. See Fig. 1:

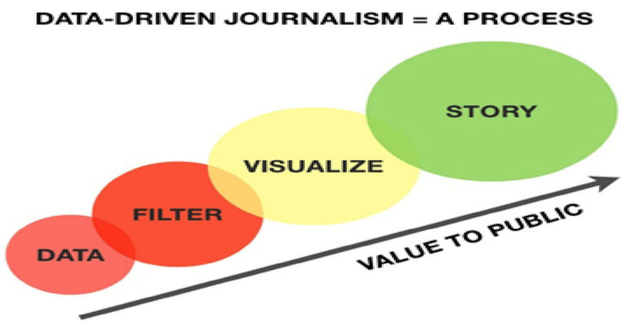


Fig. 1 The process chart for visualizing data mining

4.2 Data driven news visualization

The news data visualization is the application of visualization technology in the field of journalism, the basic principle is universal, the difference is that news, the data integration together with news value, to spread the news and meet audience demand for access to information. The application field of news visualization technology, it involves not only classical computer graphics technology, image-processing technology and computer audio technology, but also more data analysis and visual processing technologies based on new media and cloud computing involved in. With these technologies, we can not only transform data into images and graphics, but also transform them into data. At present, visualization has become the most important form of communication data in the news. With the help of scientific visualization, data visualization and information visualization and knowledge visualization of the four aspects of content, can realize the effective combination of graphics theory and image technology, let a large amount of mathematical and statistical data displayed in the form of graphics and images. It can said that the data visualization with news data can expand the news category of space-time, deepen people cognition on the news event, found that the traditional news cannot reflect the logic, to maximize the deep connotation of news reports. After the completion of visualization work, data news can use the social media platform to allow readers to personalize their applications according to their interests and needs, and achieve more positive news value [23].

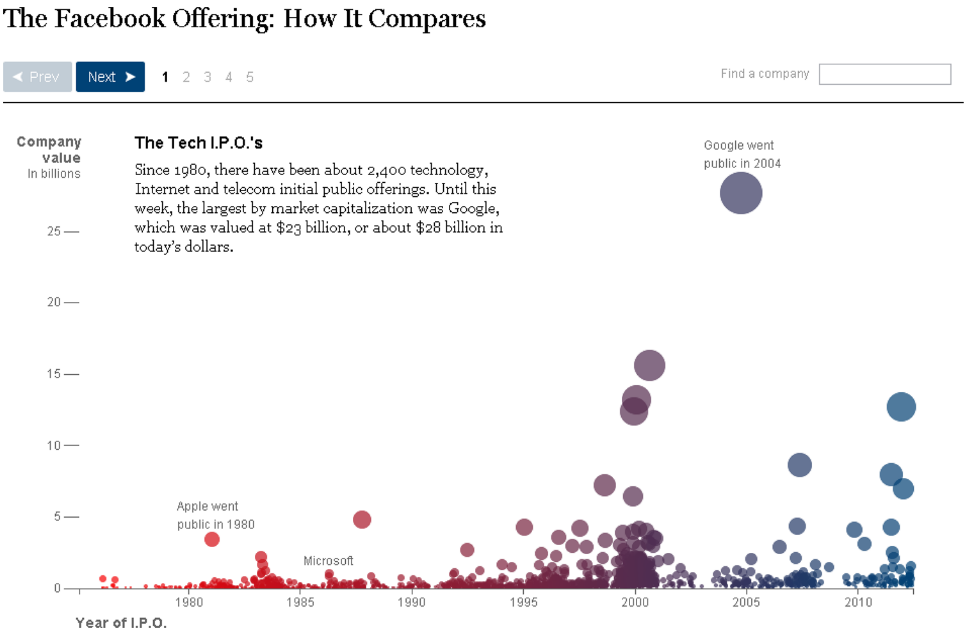
4.3 The main display mode of data news visualization

The main ways of data news visualization are static information graph, dynamic information graph and interactive graph. Static charts can implemented by statistical software. Dynamic information graph and interactive graphs are completely dependent on the internet; Interactive diagrams designed by editors, designers, and programmers to visualize data on a web page by programming, through the computer or mobile phone transmission. Users can increase or screen the content of display or visual analysis by operating computers or mobile phones. This model transforms boring data into vivid charts, and transforms complex concepts into understandable video and sound, thus enhancing the user centered visual cognition and interactive experience.

4.3.1 Static information graph

Static information graph is one of the most commonly used visual representations in data news reports. The reporter hopes to interpret more complex information with such images that contain data and information [24]. The chart

Fig. 2 The Facebook offering: how it compares



The Facebook Offering: How It Compares

The Tech I.P.O.'s

Since 1980, there have been about 2,400 technology, Internet and telecom initial public offerings. Until this week, the largest by market capitalization was Google, which was valued at $23 billion, or about $28 billion in today's dollars.

mainly transmits the news data and the information through the chart, the graph, the graph, the form, the map, the animation, the video and so on the visual chemical tool.

4.3.2 *Dynamic information graph*

The application of dynamic charts in data news is to make the abstract data meaning more graphically expressed in the form of rich visual charts, and to reflect the trends of these data development in a dynamic form. In May 2012, Facebook, the world's largest social networking site, landed on NASDAQ. What Facebook IPO means? Will it be a good choice to invest in new technology companies? The data visualization team of the New York Times better answered these questions with their works "The Facebook Offering: How It Compares" [25]. See Fig. 2.

In this work, there are five dynamic bubble charts, which represent the size of a company's prospectus or stock ups and downs. These five bubble charts review the development of technology companies in the past few decades, compare Facebook with more than 2000 technology companies to help users understand the prospects of Facebook IPO by recognizing the general development of technology companies. This work is an excellent data news report, to some extent, is also a successful social science research, so that people in recent decades the development of science and technology enterprises in the United States at a glance.

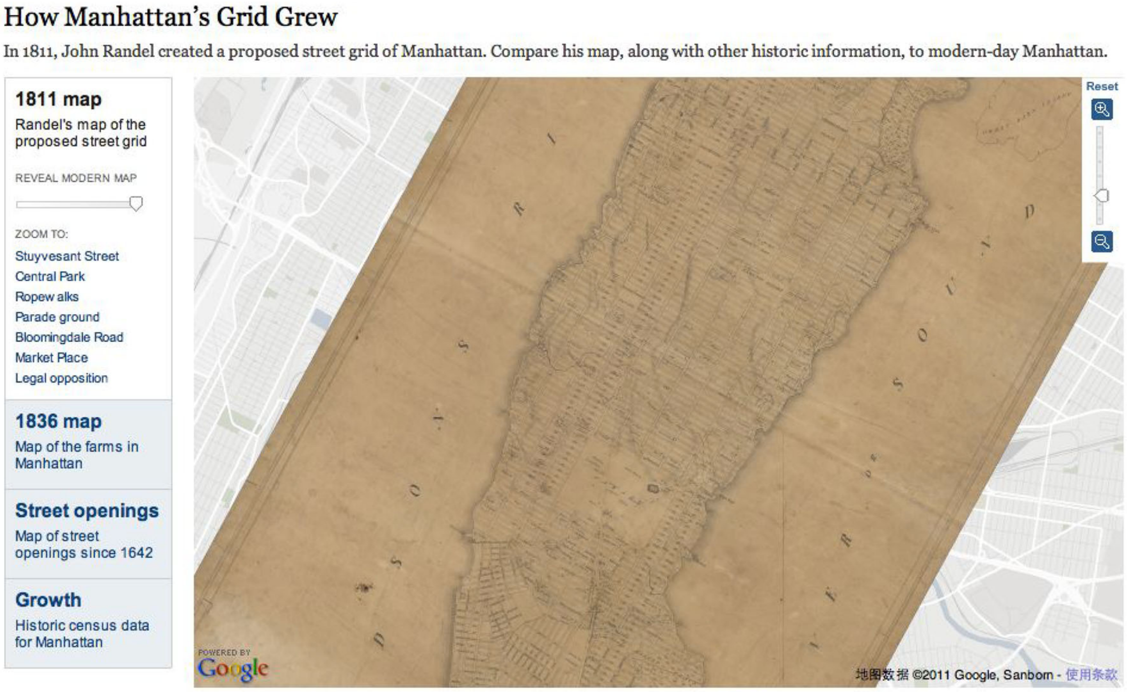
4.3.3 *Dynamic interactive graph*

In specific reports, spatial geographic data itself is the main body of news. Such as the changes of urban, space development, the scope of pollution events, the occurrence of earthquake aftershocks and so on. Writing in words is hard to convey the sense of space and wholeness that the news wants; in this case, building a dynamic interactive graph can solve the problem. This approach allows the reader to manipulate the information icon by itself, so that it is easy to understand the content of the information as a whole.

In 2011 to commemorate the birth of the map of New York on the occasion of 200 years, "New York Times" planning and producing interactive map works of "How Manhattan's grid grew" to show changes in street map of Manhattan in 200 years, see Fig. 3. Drag the map of the time bar, the user can clearly observe the 200 years of gradual changes in the road in Manhattan. In this work, the reader can clearly see how Manhattan has developed from the original minimal origin to the present. At the same time, this work also shows the visible data of population density in different regions of Manhattan in different periods. The degree of color density used to show the degree of the population density. The development of human beings and the development of Manhattan skillfully combined.

5 Analysis of data mining and visual production of data news

Different from the general news reporting based on text narration; data news centered on data and closely organizes reports around the data. At the same time, various technologies related to data have given an important place in news production. The production process of data news is getting data—analysis data—visualizing data. In this process, the data analyzed and visualized to form a wonderful news story,



How Manhattan's Grid Grew

In 1811, John Randel created a proposed street grid of Manhattan. Compare his map, along with other historic information, to modern-day Manhattan.

Fig. 3 How Manhattan's grid grew

which reflects the public value of the data. We take "urban comprehensive strength" as an example to illustrate the production process of data news.

5.1 Research background

Today's world has entered the era of global economic integration. The city, as the country's economic, political, technological and educational cultural development center, has become the protagonist of the economic cycle, the main factors that determine the status, function and future development of each city are the comprehensive economic strength possessed by each city. Urban comprehensive strength refers to a city in a certain period in the economy, society, infrastructure, environment, science and technology, culture and education and other areas have the actual strength and development ability of the set. How to measure the development status of a city through these indicators has become a problem that data journalists need to think.

5.2 Data sources

"Main Indicators of Provincial Capitals and Cities Specially Designated in the State Plan (2015)", Data come from China statistical yearbook 2016.

5.3 Research purpose

Purpose 1 Through the data analysis, report the internal relations of each industry index of the city comprehensive economic strength.

Purpose 2 Through data analysis, report the main influencing factors of urban comprehensive economic strength.

Purpose 3 Through the data analysis, report the difference of comprehensive economic strength of the city.

5.4 Research method

5.4.1 Factor analysis

Factor analysis is a statistical technique that studies the extraction of common factors from variable groups. The basic starting point of factor analysis is that the original index integrated into fewer indicators. These indicators can reflect most of the information of the original indicators (variance), there is no correlation between these comprehensive indicators. The mathematical model obtained as follows:

$$\begin{aligned} W_1 &= \kappa_{11}F_1 + \kappa_{12}F_2 + \cdots + \kappa_{1\mathrm{i}}F_i + E_1 \\ W_2 &= \kappa_{21}F_1 + \kappa_{22}F_2 + \cdots + \kappa_{2\mathrm{i}}F_i + E_2 \\ &\quad \ldots\ldots \\ W_n &= \kappa_{n1}F_1 + \kappa_{n2}F_2 + \cdots + \kappa_{n\mathrm{i}}F_i + E_n. \end{aligned} \tag{1}$$

The above models summarized and the results are obtained:

$$W_n = \kappa_{j1}F_1 + \kappa_{j2}F_2 + \cdots + \kappa_{ji}F_i + E_n, \tag{2}$$

where κ represents the standardized fraction of the n-th variable. F represents the common factor between variables. i represents the number of all variables studied in this paper. E represents the only variable that exists for the variable W. κ represents the amount of load for the variable. W represents the contribution of the m-th factor to the variance of the n variables.

5.4.2 Regression analysis

Regression analysis is a statistical analysis method to determine the quantitative relationship between two or more than two variables. According to the number of variables involved, the regression analysis divided into one element regression and multiple regression analysis. According to the number of dependent variables, the regression analysis can divided into simple regression analysis and multiple regression analysis. The mathematical model obtained as follows:

Assuming that the presence of the event A is represented by an independent variable, and then the main range of the event A variable value is represented as $[-\infty, +\infty]$. In the factor analysis model, when the variable x satisfies the relational representation of $x > \alpha$, the event A will occur, that is, $y = 1$. Otherwise, the event will not happen, $y = 0$.

The linear relationship between the reaction variable and the independent variable indicated as follows:

$$y_m^* = c + dx_m + \varepsilon_m, \tag{3}$$

Among them, y_{m}^* is the reaction variable of the actual observation value of the event, and x_i represents the independent variable, that is, the explanatory variable. C stands for regression intercept and constant term. ε means regression coefficients. E represents the error term and conforms to the standard normal distribution.

By converting the formula (1), the formula is as follows:

$$\begin{aligned} P(y_m = 1/x_m) &= P[(c + dx_m + \varepsilon_m) >] \\ &= P[\varepsilon_m > (-c - dx_m)]. \end{aligned} \tag{4}$$

Among them, P is the probability of occurrence of events, and the range is expressed as [0, 1]. When $0 < \mathrm{P} < 1$, the event exists. Otherwise, with the increase of P, the probability of occurrence of events is more and more high.

Taking into account that the variable value needs less than a particular value, the formula is given by using the distribution relation of Logistic distribution and normal distribution:

$$P(y_m = 1/x_m) = P[\varepsilon_m \leq (c + dx_m)] = F(c + dx_m). \tag{5}$$

Among them, F is the cumulative distribution function.

Assuming that the mean value of the Logistic distribution is 0 and the variance is 1, the results can be obtained:

$$P(y_1 = 1/x_m) = P[\varepsilon_m \leq (c + dx_m)] = \frac{1}{1 + e^{-\varepsilon_m}}. \tag{6}$$

In formula (4), $P \in [-1, +1]$ can obtained regardless of any value taken by ε_m.

The conditional probability of occurrence is $P(y_m = 1/x_m) = p_{\mathrm{m}}$, and the conditional probability of no event is

$$1 - p_m = 1 - \left(\frac{e^{c-dx_m}}{1 + e^{c+dx_m}}\right) = \frac{1}{1 + e^{c+dx_m}}. \tag{7}$$

B represents the ratio of the occurrence of an event to the event that does not occur:

$$\frac{p_m}{1 - p_m} = e^{c+dx_m}. \tag{8}$$

For event ratio B, take the logarithm as:

$$\ln\left(\frac{p_m}{1 - p_m}\right) = c + dx_m. \tag{9}$$

When B is from 1 to 0, the greater the value is, the greater the probability of an event is.

6 Data analysis

6.1 The internal relations of various industrial indexes of urban comprehensive economic strength

This paper makes a simple correlation analysis of the three components of the regional GDP: the benefit of the first industry, the benefit of the second industry and the benefit of the third industry (Table 1).

Correlation analysis shows that the three components of regional GDP. There is a strong correlation between the “second industry” and “the third industry”. In addition, the significant level of 0.01 was significant, and the correlation between the other variables was not significant.

Taking “gross domestic product” as dependent variable, taking “population at the end of the year” and the other variables as independent variables, multiple linear regression was carried out (Tables 2, 3).

The result shows the goodness of fit of the model is increasing in turn, and the final model adjusted R^2 为 0.0987, the goodness of fit of the model is very well (Tables 4, 5, 6).

Through multiple linear regression analysis, we can find that there is a significant relationship between the regional GDP and total retail sales of consumer goods, total value of import and export, investment in fixed assets, total population in China. The relationship with other variables is not significant. The total retail sales of consumer goods, total value of import and export, investment in fixed assets play a positive role in regional GDP; The total population has a reverse effect on regional GDP.

Table 1 Correlation

	Primary industry	Secondary industry	Tertiary industry
Primary industry			
Pearson correlation	1	0.347*	0.078
Sig. (2-tailed)		0.038	0.651
Sum of squares and cross-products	1,710,352.502	5,810,533.038	2,552,443.628
Covariance	48,867.214	1,66,015.230	72,926.961
N	36	36	36
Secondary industry			
Pearson correlation	0.347*	1	0.795**
Sig. (2-tailed)	0.038		0.000
Sum of squares and cross-products	5,810,533.038	163,775,278.850	254,078,398.030
Covariance	166,015.230	4,679,293.681	7,259,382.801
N	36	36	36
Tertiary industry			
Pearson correlation	0.078	0.795**	1
Sig. (2-tailed)	0.651	0.000	
Sum of squares and cross-products	2,552,443.628	254,078,398.030	623888092.250
Covariance	72,926.961	7,259,382.801	17,825,374.064
N	36	36	36

* Correlation is significant at the 0.05 level (2-tailed)
** Correlation is significant at the 0.01 level (2-tailed)

Table 2 Variables entered/removed

Model	Variables entered	Variables removed	Method
1	Total retail sales of consumer goods	.	Stepwise (criteria: probability-of-F-to-enter ≤ 0.050, probability-of-F-to-remove ≥ 0.100).
2	Total value of import and export	.	Stepwise (criteria: probability-of-F-to-enter ≤ 0.050, probability-of-F-to-remove ≥ 0.100).
3	Investment in fixed assets	.	Stepwise (criteria: probability-of-F-to-enter ≤ 0.050, probability-of-F-to-remove ≥ 0.100).
4	Total population	.	Stepwise (criteria: probability-of-F-to-enter ≤ 0.050, probability-of-F-to-remove ≥ 0.100).

Dependent variable: gross regional product

Table 3 Model summary

Model	R	R^2	Adjusted R^2	Std. error of the estimate
1	0.966[a]	0.933	0.931	1603.844
2	0.983[b]	0.966	0.964	1161.246
3	0.992[c]	0.985	0.983	790.530
4	0.994[d]	0.988	0.987	703.083

[a] Predictors: (constant), total retail sales of consumer goods

[b] Predictors: (constant), total retail sales of consumer goods, total value of import and export

[c] Predictors: (constant), total retail sales of consumer goods, total value of import and export, investment in fixed assets

[d] Predictors: (constant), total retail sales of consumer goods, total value of import and export, investment in fixed assets, total population

Table 4 ANOVA

Model	Sum of squares	df	Mean square	F	Sig.
1					
Regression	1, 173, 900, 569.592	1	1, 173, 900, 569.592	456.360	0.000[a]
Residual	84, 886, 395.867	33	2, 572, 315.026		
Total	1, 258, 786, 965.459	34			
2					
Regression	1, 215, 635, 243.182	2	607, 817, 621.591	450.739	0.000[b]
Residual	43, 151, 722.277	32	1, 348, 491.321		
Total	1, 258, 786, 965.459	34			
3					
Regression	1, 239, 413, 915.332	3	413, 137, 971.777	661.087	0.000[c]
Residual	19, 373, 050.127	31	624, 937.101		
Total	1, 258, 786, 965.459	34			
4					
Regression	1, 243, 957, 186.355	4	310, 989, 296.589	629.118	0.000[d]
Residual	14, 829, 779.104	30	494, 325.970		
Total	1, 258, 786, 965.459	34			

Dependent variable: gross regional product

[a] Predictors: (constant), total retail sales of consumer goods

[b] Predictors: (constant), total retail sales of consumer goods, total value of import and export

[c] Predictors: (constant), total retail sales of consumer goods, total value of import and export, investment in fixed assets

[d] Predictors: (constant), total retail sales of consumer goods, total value of import and export, investment in fixed assets, total population

Table 5 Coefficients

Model	Unstandardized coefficients		Standardized coefficients beta	t	Sig.
	B	Std. error			
1					
(Constant)	– 382.141	474.263		– 0.806	0.426
Total retail sales of consumer goods	0.000	0.000	0.966	21.363	0.000
2					
(Constant)	323.957	366.091		0.885	0.383
Total retail sales of consumer goods	0.000	0.000	0.773	16.197	0.000
Total value of import and export	0.000	0.000	0.265	5.563	0.000
3					
(Constant)	– 268.037	267.060		– 1.004	0.323
Total retail sales of consumer goods	0.000	0.000	0.568	12.258	0.000
Total value of import and export	0.000	0.000	0.367	10.078	0.000
Investment in fixed assets	0.406	0.066	0.202	6.168	0.000
4					
(Constant)	– 363.305	239.588		– 1.516	0.140
Total retail sales of consumer goods	0.000	0.000	0.592	14.104	0.000
Total value of import and export	0.000	0.000	0.358	11.028	0.000
Investment in fixed assets	0.586	0.083	0.291	7.030	0.000
Total population	– 1.294	0.427	– 0.119	– 3.032	0.005

Dependent variable: gross regional product

Table 6 Coefficients

Model	Sig.
1	
(Constant)	0.426
Total retail sales of consumer goods	0.000
2	
(Constant)	0.383
Total retail sales of consumer goods	0.000
Total value of import and export	0.000
3	
(Constant)	0.323
Total retail sales of consumer goods	0.000
Total value of import and export	0.000
Investment in fixed assets	0.000
4	
(Constant)	0.140
Total retail sales of consumer goods	0.000
Total value of import and export	0.000
Investment in fixed assets	0.000
Total population	0.005

Dependent variable: gross regional product

Table 7 KMO and Bartlett's test

Kaiser–Meyer–Olkin measure of sampling adequacy	.816
Bartlett's test of sphericity	
Approx. Chi-Square	1107.074
df	120
Sig.	0.000

6.2 The main influent factors of various industrial indexes of urban comprehensive economic strength

The common factors extracted from each variable that constitutes the comprehensive economic strength of the city (Fig. 4; Table 7).

Through KMO and Bartlett's test, the model's Sig.is 0.000, It shows that these variables are very suitable for factor analysis (Table 8).

Most of the variance of the common factor is more than 80%, so the extraction of the common factors can explain the variables (Table 9).

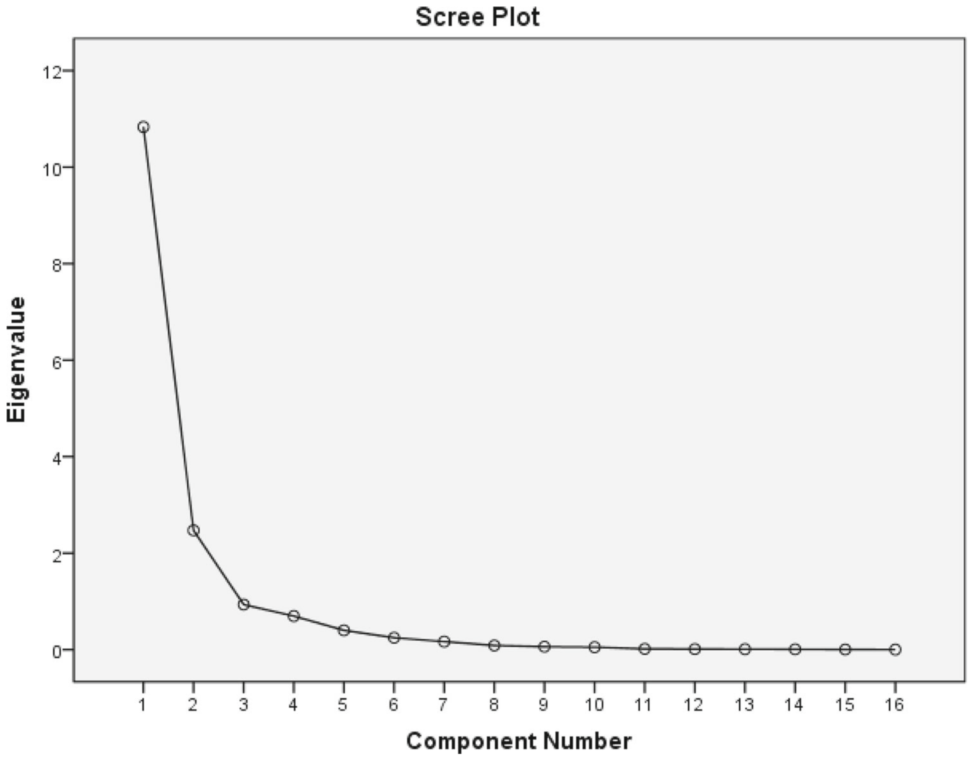


Fig. 4 The scree plot of various industrial indexes of urban comprehensive economic strength

Table 8 Communalities

	Initial	Extraction
Total population	1.000	0.947
Gross regional product	1.000	0.955
General public budget revenue	1.000	0.935
General public budget expenditure	1.000	0.945
Investment in fixed assets	1.000	0.840
Balance of savings deposit of urban and rural residents at year end	1.000	0.932
Average wage of staff and workers	1.000	0.822
Postal offices at year end	1.000	0.773
Subscribers of fixed telephones at year end	1.000	0.945
Total retail sales of consumer goods	1.000	0.906
Total value of import and export	1.000	0.929
Public vehicles under operation	1.000	0.805
Total enrollment of regular institutions of higher education	1.000	0.412
Hospitals and health centers	1.000	0.914
Licensed (assistant) Doctors	1.000	0.835
Total volume of industrial waste water discharged	1.000	0.411

Extraction method: principal component analysis

Table 9 Total variance explained

	Initial eigenvalues			Extraction sums of squared loadings			Rotation sums of squared loadings		
	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %
1	10.834	67.715	67.715	10.834	67.715	67.715	8.135	50.845	50.845
2	2.472	15.453	83.168	2.472	15.453	83.168	5.172	32.322	83.168
3	0.935	5.846	89.013						
4	0.696	4.350	93.363						
5	0.400	2.502	95.865						
6	0.247	1.542	97.407						
7	0.167	1.041	98.448						
8	0.087	0.543	98.991						
9	0.061	0.380	99.370						
10	0.051	0.321	99.692						
11	0.016	0.101	99.792						
12	0.012	0.074	99.867						
13	0.010	0.061	99.928						
14	0.007	0.042	99.970						
15	0.003	0.019	99.989						
16	0.002	0.011	100.000						

Extraction method: principal component analysis

The initial eigenvalues of two common factors are greater than 1, so we extract two common factors. The cumulative variance contribution rate of the two common factors reached 83%, almost covering all the information of the original variable (Table 10).

6.3 The rank of various industrial indexes of urban comprehensive economic strength

After factor analysis, the analysis shows that the overall economic strength of all cities is ranked as follows: Shanghai,

Table 10 Component matrix

	Component	
	1	2
Total population	0.723	0.652
Gross regional product	0.968	– 0.137
General public budget revenue	0.929	–.270
General public budget expenditure	0.958	–.165
Investment in fixed assets	0.711	.578
Balance of savings deposit of urban and rural residents at year end	0.955	– 0.139
Average wage of staff and workers	0.783	– 0.458
Postal offices at year end	0.776	0.412
Subscribers of fixed telephones at year end	0.955	– 0.184
Total retail sales of consumer goods	0.950	– 0.047
Total value of import and export	0.789	– 0.554
Public vehicles under operation	0.802	– 0.401
Total enrollment of regular institutions of higher education	0.410	0.494
Hospitals and health centers	0.698	0.654
Licensed (assistant) Doctors	0.903	0.139
Total volume of industrial waste water discharged	0.640	0.038

Extraction method: principal component analysis
Two components extracted

Beijing, Shenzhen, Guangzhou, Tianjin. Chongqing, Hangzhou, Chengdu, Nanjing, Ningbo, Wuhan, Qingdao, Dalian, Changsha, Xi'an, Shenyang, Xiamen, Zhengzhou, Jinan, Fuzhou, Harbin, Hefei, Shijiazhuang, Changchun, Kunming, Urumqi, Nanning, Guiyang, Taiyuan, Nanchang, Yinchuan, Lanzhou, Xining, Hohhot, Haikou, Lhasa. See Figs. 5 and 6.

6.4 Research conclusion

Through the above research, we can understand China's urban comprehensive economic strength from a macro perspective, which has important reference and guiding significance for the development of China's cities in the future. For example, according to the conclusion of the regression analysis, in order to improve regional GDP, China's cities must actively strengthen the development of consumer goods retail, import and export, and fixed assets

Also, the factor analysis the subsequent analysis showed that in the top of the eastern city, in the post is basically a Midwestern city, because the city economy often represents an area of advanced productive forces, so as to make the balanced development of economy in China, strengthen the construction of the Midwest is very necessary.

7 Recommendations and suggestion

We cannot just be happy with the positive impact of data news. It is necessary to point out that big data is a double-edged sword: if we can reasonable control, proper use, so it is beneficial to the journalists to expand horizons and thinking dimensions; if the improper disposal will make journalists lose sight, not to make the correct judgment, bogged down in the dilemma inextricably.

7.1 If lack of correct judgment criteria and value judgment standards, communicators and audiences will misled by big data

Data comes from every aspect of society. The data provider can be either an authority or a department, a professional or other social institutions and members of society. Different subjects have different motivations, responsibilities, seriousness and accuracy in providing data. There are also differences in credibility between data: both higher reliability data and false and wrong information. Both unintentional error data, but also malicious fake bad data. In the huge amount of information, if the specific subject does not make the corresponding choice, it will become the slave of information"; if there is a lack of judgment about the value standard, it may mislead the audience by spreading the information that is biased. In this sense, big data puts forward new and higher requirements and quality requirements for journalists [26].

Fig. 5 The classification of various industrial indexes of urban comprehensive economic strength

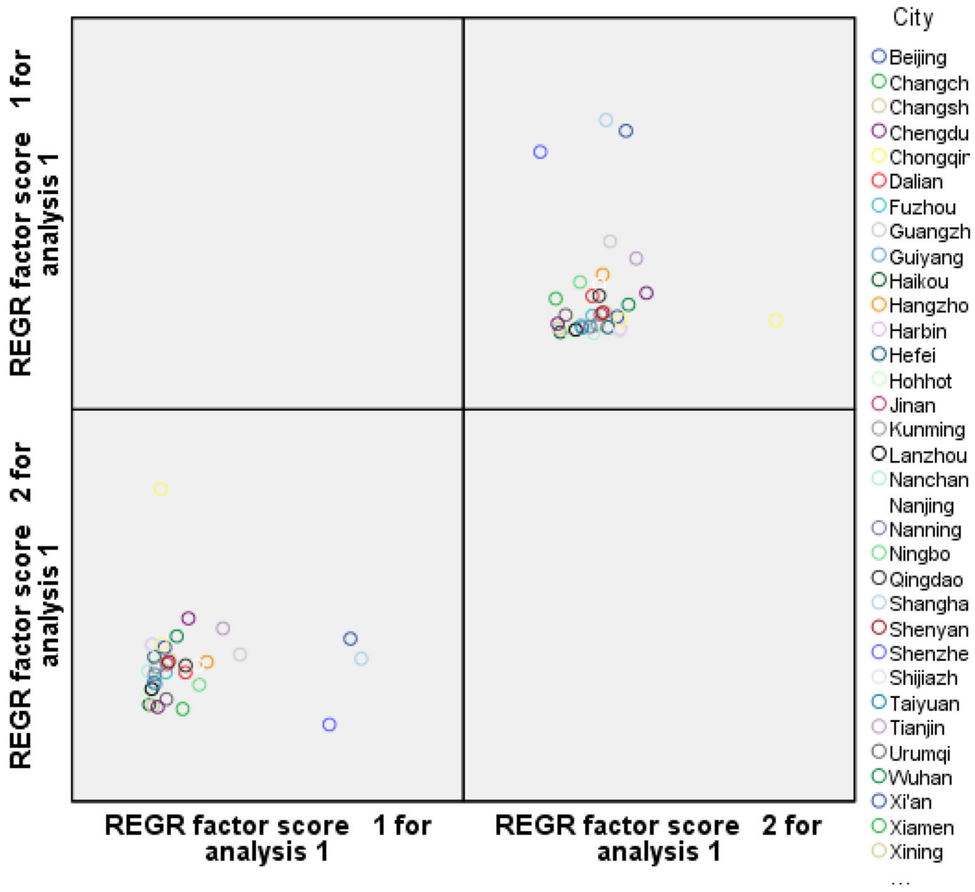
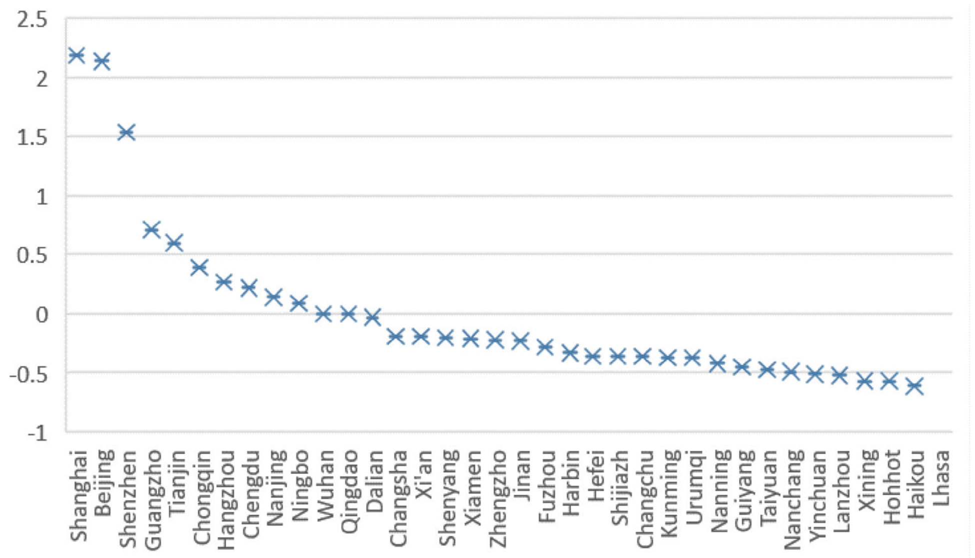


Fig. 6 The rank of various industrial indexes of urban comprehensive economic strength



7.2 Excessive superstition and dependence on big data will make the indispensable qualitative analysis absent

The object of news report is all kinds of facts, which have news value in social life. Some facts can digitized, some facts can digitized, and some facts cannot processed and expressed by numbers and digitization. Therefore, we cannot advocate using data for any type of news without indiscriminately, and not limit the data report as the direction of news reform and the trend of news development. Improving the ability of big data processing may be one of the important ways to enhance the competitiveness of the news media in the future, but it does not seem to be the only content and the whole content of the core competitiveness of the media.

7.3 News composed of data entirely is not good at telling lively and meaningful stories

News reports should be objective reports. However, it does not exclude that some reports can and should be written with human interest and interest. News value is one of the core concepts in western journalism theory, and one of the elements in news value is interesting. Data is lack of story elements in the strict sense. It is not appropriate to say that data is not interesting, but it is undeniable that a large number of data will not be interesting.

8 Development proposals

Based on the above research, this paper argues that in the background of large data; need to optimize the data visualization path of news further, by increasing the interaction experience, and improving the quality of news value and news topic.

8.1 Improve interactive experience

Under the background of big data, in order to improve the level of data visualization narration of news, news media should be through a variety of ways to strengthen the news itself interactive experience, with micro-blog, WeChat and other social networking platform, let more audience to participate. In order to achieve this, the news media need to use interactive graphics, dynamic graphics and video clips in the process of visual narrative, to meet the audience's sense of participation constantly.

8.2 Choose the topic with news value

In the era of big data, the data source of news media is very rich, so, in order to enhance the efficiency and effectiveness of data visualization, news and data should treated equally on the perspective of importance. For example, in the event of national concern, the news media should use animation data map to show the various changes of events. At the same time, from the perspective of proximity, through the big data analysis and screening, let the audience can find the link between the data and the news, let news visualization more convincing and influence. In terms of timeliness, the data news media should constantly strengthen the value of the topic, in order to complete the visual narrative of data news the first time, so that the effectiveness of news dissemination improved significantly [27].

8.3 Improve the quality of data news

The experience and lessons show that optimize the transmission mode of data news not only need to further enhance the quality of the data news, also need to complete the collection of feedback of the audience on the news through the big data technology. In this way, the dissemination effect of data news visualization will significantly improve. At the same time, the news media should further build the news work team, and create talents and intelligence conditions for the visual narrative of data news. Through the capability cultivation of data mining, information analysis and resource integration, improve the impact of data news visualization constantly [28].

References

1. Wen, W., Li, B.: Data news reports in the big data era: taking the Guardian newspaper as an example. Mod. Commun. **35**(5), 139–142 (2013)
2. Steve, Lohr.: The age of big data. New York Times, New York (11 Feb 2012)
3. Li, D., Wang, S.: On spatial data mining and knowledge discovery (SDMKD). Geomat. Inf. Sci. Wuhan Univ. **26**(6), 491–499 (2001)
4. Li, D.R., Cheng, T.K.D.G.: Knowledge Discovery from Gishie Canadian Conference on GIS, Ottawa, pp. 1001–1012 (1994)
5. Zhao, D.: Data mining: principles, methods and application. Mod. Libr. Inf. Technol. **16**(6), 41–44 (2000)
6. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., et al.: Advances in Knowledge Discovery and Data Mining, p. 18. AAAI Press, Cambridge (1996)
7. Witten, I.H.: Data Mining—Practical Machine Learning Tools and Techniques, 2nd edn. Machinery Industry Press, South Norwalk (2005)
8. Chen, M.S., Han, J., Yu, P.S.: Data mining: an overview from database perspective. IEEE Trans. Knowl. Data Eng. **8**, 866–883 (1996)
9. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Academic Press, San Francisco (2001)
10. Graham Rowe, D.: Big Data: The Next Google. Nature, UK. http://www.Nature.com/news/2008/080903/full/455008a.html (2008)
11. Jiang, X., Zhou, D.: A new data mining processing model. Comput. Mod. **2**, 18–20 (2003)
12. Wang, L.: The summarization of present situation of data mining research. Books Inf. **5**, 41–46 (2008)
13. Lewis, S.C.: Journalism in an era of big data. Digit. J. **3**, 321–330 (2015)
14. Parasie, S., Dagiral, E.: Data-driven journalism and the public good: "computer-assisted-reporters" and "programmer-journalists" in Chicago. New Media Soc. **15**(6), 853–871 (2013)
15. Wang, K.: Visualization of sports journalism in the era of big data: advantages and challenges. Sports Res. Educ. **31**(1), 13–17 (2016)
16. Anderson, C.W.: Between the unique and the pattern: historical tensions in our understanding of quantitative journalism. Digit. J. **3**, 349–363 (2014)
17. Lang, J., Yang, H.: Data news: the innovation path of news visualization communication in the big data era. Mod. Commun. **3**, 32–36 (2014)
18. Gray, J., Chambers, L., Bounegru, L.: The Data Journalism Handbook. O'Reilly Media, Sebastopol (2012)

19. Fang, J., Yan, W.: Data news from a global perspective: philosophy and practice. Int. Press **35**(6), 73–83 (2013)
20. Rodgers, Y.: Data News Trend: Releasing the Power of Visual Report: Facts are Sacred: The Power of Data. Renmin University of China press, Beijing (2015)
21. Dove, G., Jones, S.: Narrative visualization: sharing insights into complex data. Interf. Hum. Comput. Interact. **1**, 21–23 (2012)
22. Rodríguez, M.T., Devezas T.: Telling Stories with Data Visualization. Workshop, pp. 7–11 (2015)
23. Erdmann, E., Boczek, K., Koppers, L., et al.: Machine learning meets data-driven journalism: boosting international understanding and transparency in news coverage. (2016)
24. Lei, Y.: Cross Media Journalism Theory and Practice, vol. 134. Renmin University of China press, Beijing (2006)
25. http://www.NYtimes.com/interactive/2012/05/17/business/dealbook/howtheFacebookofferingcompares
26. Baiquan, D.: Data news: value and limitation. **7**, 6–10 (2014)
27. Chen, S.: Visualization model and path optimization of data news visualization in big data background. Publishing Wide-Angle, **10**, 62–64 (2017)
28. Tabary, C., Provost, A.M., Trottier, A.: Data journalism's actors, practices and skills: a case study from Quebec. J. Theory Pract. Crit. **17**(1), 41–47 (2016)

Erna Qi is a postdoc of the school of management of Huazhong University of Science and Technology. Her research areas: Media communication, New Media, Brand Communication.

Xingrui Yang is a postdoc of the school of management of Huazhong University of Science and Technology. Her research areas: Enterprise management, Technology Innovation.

Zongjun Wang is a professor of management and dean of the school of management of Huazhong University of Science and Technology. His research areas: Enterprise management, Technology Innovation.