2022 2nd International Conference on Networking, Communications and Information Technology (NetCIT) | 978-1-6654-9273-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/NETCIT57419.2022.00119

Analysis of Data News Production Mode Based on LSTM

Ran Zhang*
Leshan Vocational and Technical College
Leshan, 614000, China
100177808@qq.com

Jing Yang
University of Electronic Science and Technology of China
Chengdu, 611731, China
49466932@qq.com

Xutao Liu
Singapore National computer science
Chengdu, 610095, China
34035771@qq.com

***Abstract*—Data news is a kind of optimized expression of news content to realize the integration of data, visualization and narrative. This paper establishes the process of data news production based on the long short term memory neural network model, and focuses on the process of long short term memory for data collection, news keyword editing, database source data fitting and matching, and conducts tentative research on the production of data news. The results show that data news can provide personalized service according to user interaction behavior and realize information optimization and data value-added. A batch of data news content for reference was produced by matching the World Bank database and combining with the h a sealth theme of this paper.**

Keywords—data news, long short term memory, information editor, model analysis

I. Background

With the rapid development of new media technology in the market, information and resources have undergone essential changes in the process of social communication. Not only has the communication environment changed with the update of technology, but also has the means of information communication changed. Data news is a new type of achievement derived from the industry under the promotion of modern technology. It refers to a new news reporting mode [1] that describes events by reflecting certain social phenomena, combined with the capture, identification, mining, scheduling, statistics, processing, analysis and visual expression of events. Compared with the conventional narrative news, the data news has the advantages of intuition and visualization, which can make the audience group more intuitively understand the reported events and content. Its birth not only provides a new direction for the construction of the news industry, but also changes the traditional production process of the news industry to some extent.

Data news has the important advantages in describing the news facts, making factual judgment and predicting the development trend of events. The audience can obtain higher value information [2] in a shorter time. Literature [3] combines social media, mobile scenes, audience characteristics and so on to create personalized and customized multiple scenes, realize the three-dimensional narrative of news content, and mobilize the audience's participation in data news. From the perspective of visual design, literature [4] discusses the visual characteristics, communication advantages, limited development factors of visual news, and how to do a good job in data visual design. By analyzing the practice and application of data news in the network field, literature [5] believes that data should be deeply integrated with content, and network news should make better use of big data to show its talents. A large number of research and reports show that data news is the general trend of the future development of the news industry. In order to ensure the authenticity of the news output data and the quality and representativeness of the news content, it is very important to make the data better express the news content, make the public better perceive and recognize the complex data, improve the data reuse rate, and realize the connection and integration of data, visualization and narrative. Therefore, the main research of this paper is to establish the process of data news production, and focus on the analysis of artificial intelligence means, to conduct a tentative research on the production of data news.

II. Production Process Of Data News

A. Data and Information Collection

The exploratory data analysis method used in this paper for information retrieval pays more attention to the authenticity of the data source. The data set used requires the true distribution of the data, so the natural text data in the data published by the World Bank is selected. Artificial neuron training method emphasizes the authenticity and visualization of the data, so that the data analysts can intuitively see the hidden rules implied in the data. The long short term memory (LSTM) neural network model used in constructing the model realizes the forgetting and memory in the sequence data using the gate structure [6]. Not only can we obtain the short-distance related information in the input data, but also, we can more accurately find the dependencies with long time intervals, so it can be well applied to the processing of text data. After finding the model suitable for the present data and training the LSTM model with a large amount of text data [7], the inter-information dependence can be more accurately captured. Through the trained model, we can also predict the classification of the new data set according to the specified text.

B. News and Information Editing and Processing Process Based on Data Mining

After completing the information collection, topic selection and classification of the data news, the data mining in the big data technology is introduced to edit the above obtained news information. In the production of data news, we first need to make decisions and judge the feasibility of the news topic selection. In this process, it is

978-1-6654-9273-7/22/$31.00 ©2022 IEEE
DOI 10.1109/NetCIT57419.2022.00119

necessary to provide an important basis for feasibility judgment through data mining technology [8]. All data news information is presented in a visual way. In this process, we should choose the basic needs of the news topic according to the actual data, for example, when taking an ecological environment news as an example, it involves more information related to population density, PM2.5 and other content. Therefore, it is necessary to integrate the relevant data and information to generate the expected data news products.

C. Data Processing

The directly obtained data is relatively scattered. In order to produce the expected data news products, it is necessary to understand the data characteristics and match the corresponding mathematical models [9]. From the perspective of natural language processing, the data is trial fitting analysis. Many common linear, nonlinear mathematical models are customized in this study. In the process of data modeling, it is necessary to do some exploratory analysis of the data, mainly to have a general understanding of the overall scale of the data. Through the limited number of iterations, the feasibility of selecting the model is evaluated, so as to provide necessary conclusions for the subsequent data pre-processing and feature engineering. In this paper, we customize all kinds of algorithms, and use the consistency index to determine the target [10]. Because of the uncertainty of early information processing, each data processing process is often a variable target function and is not unique [11]. The correlation of all subject words can be obtained by modeling, so it is meaningful to conduct further analysis and information collation. At the same time, it is inevitable that the subject keywords of each document cannot converge to get the fitted model after multiple iterations of multiple algorithms and multiple models. Such random sets of subject words can be considered invalid and should abort the analysis and give feedback to the information processing center.

D. Production Details

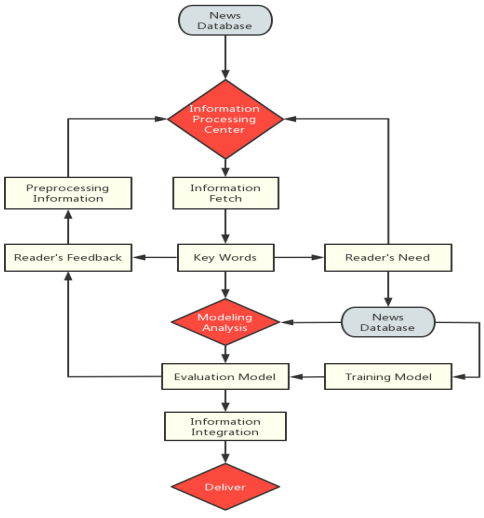

Fig. 1. Flow chart of data news production

According to the set attention content, the information processing center performs dynamic retrieval in the news database.The information processing center lists encoding and retrieval results through information fetch and key words retrieval. In this process, there will be many iterations in the information processing center through reader information feedback and reader demand acquisition. The iteration continues until the combination meets the basic requirements of network neural model modeling. In the process of combination of acquired information data and sample training, the same multiple iterations are made with the help of news database. Until the neural network model passes the evaluation and has a high consistency with the reader's concern feedback, information integration is performed. Finally, information deliver is performed to complete the entire data news production. Flow chart of data news production is shown in Fig. 1.

III. The LSTM Neural Network Model

The LSTM network model is a supervised learning model, and the LSTM network model adds three memory modules: input gate, output gate and forgetting gate, compared to the traditional RNN. This LSTM neural network model is a supervised learning model involving multiple parameters, the combination of different parameters also different results, tuning includes multiple aspects, tuning optimization is the key to model construction, if the model structure is too complex will appear the problem of overfitting, so select hidden layers of number should not be too high. This paper mainly determines the most suitable number of hidden layers through the experimental results of different hidden layer numbers. Similarly, if the number of hidden layer neurons is too easy to appear "overfitting" phenomenon, too little to accurately extract data features. The supervised learning model of multiple parameters does not have explicit rules on how many number of layers and nodes should be selected. In this paper, the pre-processing of information is added to the process node to optimize the existing information sources according to the feedback of the current hot readers to avoid repeated iteration.

In the construction of neural networks, the optimization algorithm will generally be used. On the one hand, the extreme value of the objective function is further optimized. Neural network algorithms generally first compute the difference between the objective function that is the predicted value and the true value, and then assign a gradient to calculate the loss function for each parameter. We want to find the parameters to minimize the objective function, and the purpose of the optimization is to update the parameters, and the neural network model can be optimized to get better results. We mainly choose the artificial fish group algorithm for optimization. The algorithm steps of artificial fish group algorithm of free foraging, tracking tracking and clustering make it have fast iteration speed, but the key is not easy to fall into local convergence.

LSTM neural networks are very suitable for processing long time series data. There are three different types of gates in the LSTM neuron module: input gate, forget gate and output gate. By adjusting the state of these gates, the information exchange between the hidden layers of the LSTM network can be controlled. The LSTM calculation formula is

$$i_t = \sigma(w_i \cdot [h_{t-1}, x_t] + b_i) \quad (1)$$

$$f_t = \sigma(w_f \cdot [h_{t-1}, x_t] + b_f) \quad (2)$$

$$c_t = f \cdot c_{t-1} + i_t \cdot \tilde{c}_t \quad (3)$$

$$\tilde{c}_t = \tanh(w_c \cdot [h_{t-1}, x_t] + b_c) \quad (4)$$

$$o_t = \sigma(w_0 \cdot [h_{t-1}, x_t] + b_0) \quad (5)$$

$$h_t = o_t \cdot \tanh(c_t) \quad (6)$$

In the formula, i_t, f_t, c_t and o_t are the forgetting gate, input gate, cell state, and output gate, respectively. w_i, w_f, w_c and w_o are the weight matrices. b_f, b_i, b_c and b_o are the corresponding bias vectors. σ is the function of sigmoid. *tanh* is the activation function. w is the corresponding weight coefficient matrix. h_t is the output at time t. x_t is the input at time t. $\tilde{c}_t$ is the candidate vector update value.

IV. Case Analysis

A. Aggregate Keywords

Keywords are words that reflect the chosen text theme or the main content of this article. In the retrieval of news and information, if there are accurate keywords, you can greatly improve the efficiency of retrieval. News is different from academic papers, which can directly give key words. In this paper, the LSTM neural network model is constructed, to automatically extract the keywords, and to discover and extract the information hidden in the news data through clustering. Data news questions can map the subject content, and analyzing the topic names can help to identify the hot topics. The BDP is used to generate word clouds as shown in the figure [12].

Data news search keywords mainly include: COVID-19, the Belt and Road, new, positive, pneumonia and other words, focusing on health topics. Due to COVID-19, health has become the focus of intense social attention, and this paper finally selected health data and news as a sample.

B. Information Retrieval and Encoding

Data news is the carrier of data visualization. Its planning, topic selection, topic extraction, data extraction and visualization jointly construct data news. Therefore, after the keyword determination, an important step is the data encoding. In this paper, the retrieved health hot words are encoded one by one, and the representative characteristics of the construction of data news are summarized. Each data news appears in one trait or more of the same characteristics are encoded only once. Through continuous comparison, discussion and combing, a total of 5,819 original statements were obtained, and 10 categories (X1-X10) were combined and fit out. Because of a sentence, a paragraph of text, a chart and other performance characteristics as a representation form of open coding data. The encoding and retrieval results are shown in Table I.

TABLE I. Encoding and Retrieval Results

No.	Categories	Numbers
X1	News Title	985
X2	STATISTICAL DIAGRAM	354
X3	Picture Presentation	245
X4	Background News	544
X5	High-Frequency Words	1124
X6	Author Tracked	437
X7	Social Platform	547
X8	Literature Track	548
X9	Real-Time Comment	358
X10	Associated Point	677

C. Modeling and Analysis

The LSTM algorithm is an algorithm that matches each other through keywords and key information extracts, based on the already retrieved categories. The algorithm is analyzed according to the correlation degree of different categories X1-X10 and the weight of the PI database. However, the conventional algorithm cannot distinguish the correlation degree of different categories X1-X10, and can only match one by one, and the iteration convergence is slow, or even convergence. This case uses both the LSTM algorithm and the routine algorithm for an iterative analysis.

Observe from Fig. 2 accuracy, in the process of iteration, LSTM algorithm in the accuracy trend graph test set of small shocks, with the increase of iterations, training loss and training accuracy has been gradually convergence, LSTM algorithm on the training set accuracy is gradually rising, loss value is gradually declining, the effect of the whole training set has a better and better trend. The model training is preliminarily judged to meet the expected effect. The prediction effect is also gradually approaching the best accuracy value. Compared with the conventional algorithm, the accuracy rate has been fluctuating and rising trend, and it still failed to reach the ideal accuracy value in the limited 400 simulation iterations.

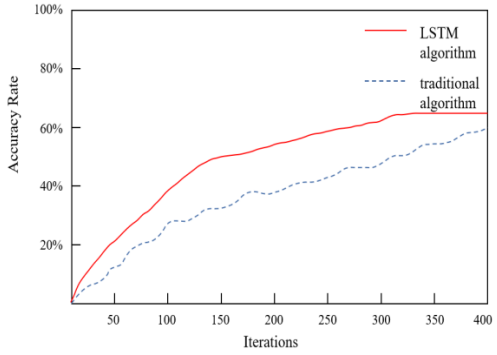

Fig. 2. Comparison of iteration times

D. Evaluation Model and Information Integration

The LSTM algorithm can mine data and present visual results, but the lack of narrative plot is not conducive to the public to interpret the data. The data news combined with information retrieval and encoding reveals the logical relationship of data hiding, makes the massive data readable through the short narrative, and helps the audience to control the data. Matching the World Bank database, combined with the health theme of this paper, produced a batch of data news content for reference. As we can know from Table II, the amount of data news generated in the massive data analysis is limited. Ten categories of keywords were obtained by fitting the original statement and optimized the keyword content by combination to match the corresponding data news.

TABLE II. EXAMPLES OF DATA NEWS

No.	Title	Derived from
1	Physical health problems of international students	X1∩X2∩X5
2	Life expectancy of female practitioners in Africa	X1∩X5∩X8
3	COVID-19 High Frequency Survey of Households and Children 2020-2021	X2∩X5
4	COVID-19 vaccination rates in China	X2∩X9
5	High COVID-19 vaccination rates are helping to restore life expectancy	X3∩X5∩X6∩X8
6	COVID-19 ages human organs by 3-4 years	X2∩X5
7	Flu drug sales in China and the United States	X2∩X4∩X6∩X10
8	Targeted prevention and control promotes orderly economic recovery	X2∩X10
9	How effective are existing vaccines in preventing infection with the Omicron variant?	X4∩X5∩X6∩X8
10	Effects of PM2.5 on life expectancy in different regions	X5∩X8
11	Regional distribution of myopia in adolescents	X5∩X8∩X9∩X10
12	Death rates of cancer patients	X5∩X9
13	Relationship between sex ratio at birth and region	X6∩X7∩X8∩X10
14	Dietary health and growth status	X7∩X9

V. CONCLUSION

Data news can provide personalized services according to user interaction behavior, and realize information optimization and data value-added. The development of the Internet has enriched the different types of information terminals. Young readers are more inclined to use their fragmented time to read high-quality data news. In data news, infographic charts are suitable for the expression of narrative spatial structure, location and details, while text is suitable for the transmission of abstract information, oral concepts and logical conditions. Joint coding of infographic charts and text can activate various cognitive schema types of readers at the same time and improve their reading effect. However, data news is currently in the exploratory stage, and the accuracy of news generation is not high, and there are still many obstacles. In the future, it is suggested to expand the construction of the sample space of data news, establish a professional data news database, and share more data resources in the government, banking, scientific research and other fields, so as to improve the accuracy and readability of data news.

REFERENCES

[1] Zhang Jing. Research on "DT Finance". Guangdong Technical Normal University, 2022.

[2] Han Mengmeng. Explore the news value behind the data in depth. China Press and Publication Wide Telegraph, 2022-09-13 (005). DOI:10.28907/n.cnki.nxwcb.2022.003152.

[3] Liu Xiaolu, Li Jie. Analysis of the "Scene" of data news under the Background of mobile Internet. Southern Media Research, 2022 (04): 34-38.

[4] Tong Ke. On Data Visual Design in News Communication. China Press Industry, 2022 (15): 120-121. DOI: 10.13854/j.cnki.cni.2022.15.001.

[5] Ma Kai, Wang Xiaonan, Han Qiang. Big data + news: boost the development of network data news new engine. Young Reporter, 2022(14): 95-96. DOI:10.15997/j.cnki.qnjz.2022.14.033.

[6] Bhavekar, Girish S., Goswami, Agam Das. A hybrid model for heart disease prediction using recurrent neural network and long short term memory. International Journal of Information Technology, 2022 (prepublish).

[7] Sun Xiaomei, Zhang Haiou, Wang Jian, Shi Chendi,Hua Dongwen, Li Juan. Ensemble streamflow forecasting based on variational mode decomposition and long short term memory. Scientific Reports, 2022, 12 (1).

[8] Agrawal, Saurabh, Sisodia, Dilip Singh, Nagwani, Naresh Kumar. Long short term memory based functional characterization model for unknown protein sequences using ensemble of shallow and deep features. Neural Computing and Applications, 2021.

[9] Li Zheng. Automatic Production Technology of Data News Based on Machine Learning Model. Wireless Communications and Mobile Computing, 2022.

[10] Jinhee Kim, Jeongsub Lim. A Study of the Characteristics and Definition of Data News in Terms of Finalists of Data Journalism Award from 2012 to 2015. Communication Theories, 2016, 12 (2).

[11] Kilburn, Faye. SunGard Adds Platts Real-Time Energy Data, News to MarketMap. Inside Market Data, 2015, 30 (21).

[12] Nasty truth about dirty data news. African Printer, 2010, 2010 (2).