New Media Data Analysis and Application: A Preliminary Analysis (Postprint)

Authors: Kaixin Wang, YAN Rui, Wang Liguo, Shuangli Wang

Date: 2023-10-08T00:00:00+00:00

Abstract

From the perspective of data analysis requirements, this paper describes the establishment of new media data collection objectives and collection methods; following an introduction to new media data preprocessing aimed at enhancing quality and efficiency, it elaborates through instantiation on the concepts, principles, and implementation methods of classical analytical methods represented by matrix analysis, correlation analysis, and regression analysis, as well as algorithmic models represented by BP neural networks; it also discusses big data analysis capabilities for massive datasets and the application design of Hadoop-based KNN classification algorithms, thereby providing theoretical foundations and technical support for new media operation enterprises to effectively address new media data analysis.

Full Text

Preamble

A Brief Analysis of New Media Data Analytics and Applications

Kai-Xin Wang[1], Rui Yan[1], Li-Guo Wang[2], Shuang-Li Wang[2*]
(1. Shanxi Media College, Jinzhong, Shanxi 030619; 2. Beihua University, Jilin, Jilin 132021)

Abstract: From the perspective of data analysis requirements, this paper describes the establishment of new media data collection objectives and collection methods. After introducing new media data preprocessing aimed at improving quality and efficiency, it elaborates conceptually and empirically on classic analytical methods including statistical analysis (represented by matrix analysis, correlation analysis, and regression analysis) and algorithmic models (represented by BP neural networks), covering their principles and implementation approaches. The paper also discusses big data analysis capabilities for massive

datasets and the application design of Hadoop-based KNN classification algorithms, providing theoretical foundations and technical support for new media enterprises to confidently address new media data analysis challenges.

Keywords: new media data analysis; BP neural network; big data analysis; KNN classification algorithm

1. Data Collection

1.1 Establishing Collection Objectives

In the current era of massive new media data volumes, collection objectives must be defined based on data analysis requirements before acquisition begins. This involves setting collection scope and eliminating redundant data to enhance data representativeness and credibility. The process requires identifying critical nodes for problem-solving from real-world issues, extracting relevant transaction feature attributes, and using these attributes to plan data analysis directions and refine collection targets.

1.2 Data Sources and Acquisition Methods

New media data are typically generated during social production, management, and operations processes, primarily sourced from online databases, social media, network public opinion, and system operation logs. New media data acquisition fundamentally involves collecting data through multiple approaches, commonly utilizing operator (or manager) databases and third-party platform data. These two methods typically involve direct acquisition from operational system servers or cloud-based collection using intelligent web crawler technology to achieve real-time data aggregation. Additionally, manual questionnaires serve as a supplement to the aforementioned methods, facilitating on-site communication between investigators and respondents and enabling precise grasp of respondents' psychological characteristics to clarify their needs.

2. Data Preprocessing

Data preprocessing refers to the processing and organization performed before primary data processing and analysis to achieve objectives such as cleaning anomalies, correcting errors, and standardizing formats. Methods including data cleaning, data integration, and data transformation improve the quality and efficiency of analysis. In the context of the new media era, external information requires not only convenient and rapid acquisition in the traditional sense but also efficient processing and analysis, precise delivery, and service operations—requirements that are crucial for nations, enterprises, and individuals alike. The deep integration of new media with modern technologies such as cloud computing, big data, and artificial intelligence provides society with

higher-quality data applications and intelligent services, offering reliable foundations for new media enterprises to identify directions, reduce costs, and plan proposals.

Data cleaning has become a common method for big data preprocessing, primarily achieving goals such as removing duplicate information, correcting existing errors, and ensuring data consistency.

3. New Media Data Analysis

New media data analysis refers to the use of appropriate analytical methods to dissect and process large volumes of new media data, making it comprehensible and reflecting the essential characteristics and inherent patterns of real-world phenomena represented by the data, thereby maximizing the utility of data. Commonly used classic analytical methods include straightforward statistical analysis, highly complex algorithmic models, and big data analysis tailored for massive datasets.

3.1 Statistical Analysis

3.1.1 Matrix Analysis Matrix analysis uses two important metrics of the data to be analyzed as horizontal and vertical axes to form four quadrants for problem analysis, proposing reasonable solutions and deriving data analysis conclusions. The Kano model, as a representative matrix analysis method, is illustrated in [Figure 1: see original paper].

In the Kano model, exciter needs represent requirements that users completely do not expect or that exist in their subconscious state, requiring mining and insight. When these needs are provided, users experience unexpected delight and demonstrate high satisfaction; if not provided, satisfaction decreases. Performance needs, as a one-dimensional factor, are directly proportional to satisfaction. These are needs in the growth stage that require attention from customers, competitors, and operating enterprises alike. They reflect competitive capability, and enterprises should focus on improving service quality for such needs. Basic needs are essential requirements for products, where service products must have relevant functions. When product functions are continuously strengthened, user satisfaction does not significantly improve, but eliminating these functions causes satisfaction to drop noticeably. Reverse factors, which users completely do not need, are inversely proportional to user satisfaction. Therefore, during product design, reverse factors should be avoided, basic needs should be well addressed, the quality of performance needs should be continuously improved, and exciter needs should be highlighted.

Each functional requirement in Kano model research has both positive and negative evaluations. Based on each functional requirement, evaluations can be formed according to four values: like, should be, acceptable, and dislike, creating a two-dimensional table to calculate Better-Worse coefficients. The Better coefficient represents the satisfaction coefficient (typically positive), while the Worse

coefficient represents the dissatisfaction coefficient (typically negative). [Figure 2: see original paper] shows the demand analysis corresponding to Better-Worse coefficients, where Product Feature 1 in the first quadrant represents the optimal performance need and can be prioritized.

Kano model construction must incorporate business characteristics; otherwise, analysis conclusions may significantly deviate from actual conditions. This requires that questionnaire design accurately reflect product features and identify appropriate survey respondents. Additionally, certain types of needs may evolve into other types over time, necessitating continuous research and product updates based on data analysis results.

3.1.2 Correlation Analysis Correlation analysis measures the degree of closeness between two or more variable factors, though correlation does not imply causation. In new media marketing, correlation strength between two products can be compared to determine whether to conduct bundled sales. Using Excel' s CORREL function, [Figure 3: see original paper] presents the correlation coefficient calculation results for Book A' s sales volume versus Books B through G during January 1-6, 2021. The correlation coefficient ranges from [-1, 1], where larger absolute values indicate stronger correlation. The results show that Books A and D have the highest correlation coefficient of 0.821326501, suggesting that bundling these two can stimulate more purchasing behavior.

This relationship can be visually presented using line charts, clearly showing that Book D' s sales volume generally increases as Book A' s sales volume increases, as illustrated in [Figure 4: see original paper].

3.1.3 Regression Analysis Regression analysis determines relationships between dependent and independent variables by establishing regression equations to express correlations and predict future changes in dependent variables. There may be multiple independent and dependent variables. The following example illustrates this application.

A transportation company sought to develop an optimized transportation plan and predict daily driver working hours based on cargo volume. Analysis revealed that daily driver working hours correlate with transportation distance and frequency. A random sample of 12 transportation activities was collected, and based on this data (shown in), a binary linear regression equation was constructed. With time as the dependent variable and distance and frequency as independent variables, Excel' s regression analysis tool estimated the regression coefficients, with results summarized in [Figure 5: see original paper].

The multiple correlation coefficient R of 0.9497 indicates high linear correlation between time and both transportation distance and frequency. Based on these results, regression coefficients can be obtained. The regression coefficient of 0.043 indicates that with frequency held constant, each additional kilometer of transportation distance increases travel time by an average of 0.043 hours.

The coefficient of 0.573 indicates that with distance held constant, each additional transportation increases time by an average of 0.544 hours. For a driver transporting 5 items along an optimal route of 150 kilometers, the regression equation y = -0.115 + 0.043 × 150 + 0.544 × 5 predicts transportation time of 9.055 hours, which shows high consistency with sampled data.

3.2 Complex Algorithmic Models

Compared to statistical analysis, more complex classic algorithmic models include decision trees, chaos theory, neural networks, ant colony algorithms, and particle swarm algorithms. These models are primarily applied in artificial intelligence, information science, control theory, machine learning, and other fields for description, analysis, prediction, optimization, decision-making, and control. These complex algorithmic models, validated repeatedly in cutting-edge science, remain applicable to new media data analysis, as demonstrated in cases such as "Research on Deep Neural Network Video New Media Short Video Personalized Recommendation Systems" [2]. The following describes the application of algorithmic models using BP neural networks as an example.

BP neural networks, fully known as error backpropagation neural networks, add hidden layers for connecting weights in multi-layer neural networks. BP neuron transfer functions can employ linear or nonlinear functions, capable of approximating any continuous function on a closed interval, enabling multiple value options for a neuron's output. BP neural networks consist of input, hidden, and output layers. Nonlinear relationships between input and output vectors can be described through nonlinear function neurons in the hidden layer. In the three-layer BP neural network structure shown in [Figure 6: see original paper], the hidden layer contains q neurons, and the output layer contains L neurons. Error backpropagation from the output layer can adjust weights between hidden and output layers. [Figure 7: see original paper] demonstrates the application of BP neural networks in forecasting network traffic load information for a service system.

Load data were measured every two hours daily, along with maximum temperature, minimum temperature, and weather characteristics (0 for sunny, 0.5 for cloudy, 1 for rainy). The data in [Figure 7: see original paper] have been normalized, comprising 15-dimensional input vectors (12 network load vectors plus 3 meteorological feature vectors) and 12-dimensional output vectors (12 network load vectors). Consequently, the network input layer has 15 neurons described by X, the output layer has 12 neurons (with logarithmic activation functions) described by Y, and the hidden layer neuron count is set to 31 (with tangent activation functions). Training the neural network Y = F(X, W) yields the weight set W, and prediction results M are calculated using M = F(K, W).

3.3 Big Data Analysis

Big data analysis processes data with massive scale, offering advantages of rapid data flow and reliability. Data analysis technology in digital new media can serve as a foundation for artificial intelligence [3]. Unlike statistical analysis and algorithmic models, big data analysis provides more complex, stable, and secure data analysis system architectures with functions including data mining, predictive analysis, and data warehousing.

In big data application systems, Hadoop serves as a distributed storage and computing framework for processing big data, widely adopted by large, medium, and small enterprises domestically and internationally [4]. Its core components include the HDFS distributed file system (scalable, highly fault-tolerant, high-performance) and MapReduce, which provides parallel computing frameworks for rule extraction from massive data [5]. Implementing KNN machine learning algorithm models under the Hadoop framework becomes more convenient and efficient.

The core idea of the KNN algorithm is that if most of a sample's K nearest neighbors in feature space belong to a particular category, the sample also belongs to that category and possesses the characteristics of samples in that category. This method determines the category of a sample to be classified based solely on the categories of its nearest neighbors. [Figure 8: see original paper] illustrates the KNN classification principle. The figure contains two types of sample data: star-shaped and elliptical, with circular points representing data to be classified. When K = 3, the three nearest neighbors to a circular point include 2 elliptical and 1 star-shaped sample; elliptical samples are in the majority, so the point is classified as elliptical. When K = 6, the six nearest neighbors include 2 elliptical and 4 star-shaped samples; star-shaped samples are in the majority, so the point is classified as star-shaped.

When training data volumes are massive, traditional KNN algorithms fail due to limited single-machine memory and computing resources. However, since each training sample is essentially unaffected by other training samples, KNN can be implemented using MapReduce. The core components of MapReduce are map and reduce functions: map divides large tasks into smaller tasks that can run simultaneously, while reduce aggregates results from multiple small tasks. For KNN, when training data volumes are large, training data can be distributed and read into map. Each input training sample in map calculates its distance to all test data and passes these to reduce, which then merges distances for the same test data, sorts and counts them to obtain class labels for test samples and outputs them. Since Hadoop' s MapReduce computing framework follows key-value pair principles, map and reduce function designs are shown in .

The above describes the application design for implementing KNN algorithms using Hadoop, achieving the goal of simple, efficient, and parallel processing for classification tasks. This method is suitable for solving critical issues in new media such as text classification, public opinion analysis, and public opinion

prediction.

Conclusion

The new media data analysis methods mentioned above each possess distinct characteristics. Statistical analysis methods require high completeness and accuracy of historical data, with simple, easily controlled analysis steps. Complex models such as BP neural networks exhibit strong learning capabilities and fault tolerance, capable of handling complex nonlinear relationships. Big data analysis provides architectural-scale solutions for massive datasets, balancing standardization with data processing flexibility, typically completed by professionals. Therefore, it is necessary to reasonably leverage their respective strengths based on actual conditions of real-world problems and resources available to new media enterprises to ensure reliable new media data analysis results.

References

[1] Duan Fengfeng. *New Media Data Analysis and Application* [M]. Beijing: Posts & Telecom Press, 2020: 100.
[2] Gao Chenfeng. Research on Deep Neural Network Video New Media Short Video Personalized Recommendation Systems [J]. *Satellite TV*, 2019(5): 16-20.
[3] Duan Rulin. Research on the Application of Data Analysis Technology in Digital New Media [J]. *Education Modernization*, 2016(36): 215-216.
[4] Yu Minghui, Zhang Liangjun. *Hadoop Big Data Development Fundamentals* [M]. Beijing: Posts & Telecom Press, 2018: 1.
[5] Liu Chunyang, Zhang Xuelong, Liu Lijun, et al. *Hadoop Big Data Development* [M]. Beijing: China Water & Power Press, 2018: 12.

Author Biographies

Kai-Xin Wang (2001-), female, from Jilin, research direction: Network and New Media.
Shuang-Li Wang (1972-), male, from Jilin City, Jilin Province, associate professor, research directions: Data Analysis and Application, Intelligent Science and Technology. (Corresponding author)

(Responsible Editor: Xiaojing Zhang)

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