Academic Education Publishing House

Industry Science and Engineering Vol. 2 No. 10, 2025

Data Analysis and Visualization of Food-Related Videos on Bilibili

Zehui Li[1], Qiuchu Li[1], QingFeng Zhou[2]

[1]College of Artificial Intelligence and Big Data, Henan University of Technology, Zhengzhou, China

[2]iFLYTEK Co., Ltd., Hefei, China

Abstract: In the current short-video content ecosystem, efficient analysis of massive data within vertical domains is crucial for content creation and platform operation. This project aims to conduct in-depth data analysis and visualization of food-related videos on Bilibili. Initially, the project utilizes HDFS for storing raw video data and employs the Spark big data processing framework for data cleaning, transformation, and statistical analysis, with the processed results imported into a MySQL database. Subsequently, the backend, built on Tomcat/Servlet and JDBC technologies, is responsible for database connection and providing structured data interfaces. The frontend adopts mainstream technologies such as Vue 3 and ECharts, retrieving data via Ajax to construct a fully functional web application. This application enables multi-dimensional visual analysis of food video content, UP Master engagement, traffic trends, and other key metrics, thereby offering data-driven insights and decision support for content creators and platform operators.

Keywords: Visualization;Spark; Vue; Bilibili Food Videos; Big Data Analytics; Servlet

1. Introduction

As a comprehensive video community highly populated by Generation Z, Bilibili has formed a distinct youth-oriented label through its unique content ecosystem. By the third quarter of 2024, Bilibili had achieved 330 million monthly active users, with users aged 18-35 accounting for over 78%. As one of the core vertical categories on the platform, the food section witnesses over 50,000 daily submissions, covering diverse content formats such as store exploration reviews and home-cooking tutorials. Similar to how in-depth analysis of customer information, service usage, and payment behaviors helps telecom operators better understand their customer base, analyzing behavioral data of food videos—including views, bullet comments, and tag correlations—empowers UP Masters and the platform to grasp content patterns, thereby formulating more effective marketing strategies and service improvement plans.

The "Bilibili Food Industry Marketing Insight Report" launched by Feigua Data (Bilibili Edition) in 2021 indicates that food-related videos excel in core metrics such as total views and bullet comment share, both reaching approximately twice the proportion of videos in the overall industry. The share of coins awarded even exceeds the industry average, further confirming the important position of the food section in Bilibili's content ecosystem [1]. Additionally, Bilibili's official "Annual Food Data Report" states that the total views of food videos surpassed 24 billion that year, a year-on-year increase of 70%. Among them, niche content such as "solo dining" and "food replication" experienced significant growth in submission volume, reflecting the diversified development trend of food content and highlighting the necessity of in-depth data analysis in this field [2].

2. Visualization Analysis Process

This project follows a typical big data analysis and visualization workflow:

(1) The core objectives include multi-dimensional statistical analysis and visualization of video data in Bilibili's food section, covering tag analysis, bullet comment interaction analysis, view trend tracking, and comparison of triple-interaction rates (likes, coins, favorites).

(2) Data preprocessing is performed prior to visualization. The raw data, sourced from the internet, includes information about Bilibili food videos such as titles, authors, release times, views, bullet comments, likes, comments, coins, and favorites. Preprocessing steps involve data cleaning (deduplication, missing value handling, outlier processing), data upload to HDFS, analysis using Spark, and import of processed results into MySQL.

(3) In the visualization module, ECharts is used to generate charts for multi-dimensional features (tags, bullet comments, interactions), including word clouds, Sankey diagrams, box plots, radar charts, and heatmaps.
(4) Finally, the visualization results are summarized and analyzed, with corresponding recommendations and future prospects proposed.
The Python-based big data analysis system has formed a mature paradigm in Bilibili video research. For example, Zhou proposed data-driven content optimization strategies through the analysis of Bilibili APP interfaces and video interaction data in emotional design research, providing references for the project's data preprocessing workflow—especially in the correlation analysis between user interaction behaviors and content features [3]. Meanwhile, Yu et al. optimized content release strategies using data analysis in their research on university libraries operating Bilibili accounts. Their conclusions on the relationship between release time and traffic provided theoretical support for the design of the view trend and time distribution analysis module [4].

3. Overview of Telecom Customer Data

3.1 Data Analysis Requirements

(1) Tag and View Analysis: Statistically analyze multi-dimensional data (views, likes, favorites, coins) of videos under different tags to reveal tag distribution patterns and potential high-value tags.
(2) Bullet Comment and Interaction Heatmap Analysis: Identify high-frequency bullet comment keywords, analyze hot topics with high interaction rates, and integrate with likes, favorites, and views to generate multi-dimensional heatmaps.
(3) Triple-Interaction Rate and Top Creator Analysis: Statistically analyze and visualize triple-interaction rates to identify high-quality content and top UP Masters.
(4) View Trend and Time Distribution: Analyze monthly view trends and the impact of release time slots on views.
(5) Word Cloud and Text Analysis: Perform word segmentation and frequency statistics on titles and tags to explore content hotspots and assist in topic selection and creation.
(6) Recommendation Provision: Help UP Masters and the platform understand content rules to develop more effective marketing strategies and service improvement plans.
Cai et al. constructed a comprehensive evaluation system including views and interaction rates in their research on internet fitness teaching video evaluation. Their method for weight assignment of multi-dimensional indicators provided methodological support for designing evaluation dimensions in tag-view analysis and triple-interaction rate analysis—especially in balancing the importance of different interaction metrics [5].

3.2 Data Source

The data is sourced from the Bilibili food video dataset on the Heywhale Community, stored in CSV format. The dataset information is shown in Table 1.

Table 1. Description of Field Types in the Raw Dataset

Field Type	Data Type	Description
Title	string	Text name of the video
Total Views	Int	Total number of times the video was watched
Total Bullet Comments	Int	Total number of bullet comments on the video
Release Time	string	Release date and time of the video
Publisher	string	Nickname of the video publisher (UP Master)
Likes	string	Total number of likes on the video
Favorites	string	Total number of favorites of the video
Comments	Int	Total number of comments on the video
Tags	string	Keyword tags associated with the video

3.3 Development Tools and Programming Languages

The frontend of the project is developed based on the Vue 3 framework, with JavaScript as the core programming language. JavaScript code implements data acquisition, processing, rendering, and interaction. Routing management uses vue-router, and state management adopts vuex—both implemented via JavaScript.
In addition, the frontend page structure is built with HTML, and styles are designed with CSS. These two languages collaborate closely with JavaScript to create an aesthetically pleasing, responsive, and interactive user interface.
In summary, the project primarily uses

JavaScript, supplemented by HTML and CSS.

Luo proposed a solution for frontend component-based development and asynchronous data loading in the design of a Vue-based responsive data visualization platform. When constructing the Vue 3 + ECharts visualization interface, this project drew on her ideas for component reuse and Ajax request optimization, effectively improving page rendering efficiency [6]. Meanwhile, Jing et al. optimized parameter configuration and interaction logic for charts such as word clouds and heatmaps in their research on ECharts data visualization. This provided technical references for generating multi-dimensional visual charts in the project—especially in the generation of word clouds for high-frequency bullet comments and the display of interaction data heatmaps, where their chart style design and data mapping methods are directly applicable [7].

4. Data Processing

Data processing is a critical step to ensure the accuracy of visualization results, mainly completed under the Spark framework with auxiliary initial cleaning using Python scripts.

4.1 Data Cleaning

(1) Garbled Text Handling: During data transmission and upload to HDFS, garbled text issues occurred due to the large amount of Chinese content. The problem was resolved by unifying the encoding format to UTF-8.

(2) Field Cleaning: Fields such as "Total Views" and "Likes" in the raw data contained extra characters. Regular expressions were used for extraction and type conversion to numeric formats.

(3) Deduplication Logic Challenges: Initially, deduplication in Spark only used title, publisher, and release time as unique identifiers. However, practical application revealed that this logic might incorrectly deduplicate videos with similar content but slightly different release times during repeated data collection, affecting data integrity.

(4) Filter Condition Optimization: The initial cleaning logic, such as filtering data with filter(col("clean_views")>0||col("clean_bullet_comments")>0), might have excessively filtered video records for certain months. Videos with zero views and bullet comments but non-zero values in other metrics (e.g., likes, favorites) would be mistakenly deleted. Subsequent targeted optimization of filter conditions ensured comprehensive data coverage.

Shi proposed a complete data cleaning process—including text garbled handling and numeric field extraction—in his research on constructing video evaluation indicator systems integrated with text data. When processing Chinese tag garbling and format conversion of "views" and "likes" fields, this project referenced his field cleaning scheme based on regular expressions, effectively improving data accuracy [8]. Additionally, the "Bilibili Video Analysis System Based on Big Data" mentioned that Spark deduplication logic should combine video MD5 values or content feature values to avoid misjudgment. This provided direction for optimizing the deduplication strategy in this project, and more dimensional unique identifiers may be introduced in the future to improve data integrity [9].

4.2 Data Segmentation

After data cleaning, Spark SQL is used to execute aggregate queries to generate indicator data required for visualization:

(1) Time Series Aggregation: Aggregate data by month or date to calculate total views, total bullet comments, etc., for trend chart generation.

(2) Tag Decomposition and Word Frequency Statistics: Split video tags, count tag occurrence frequency and average views under each tag for word cloud and box plot generation.

(3) UP Master Performance Calculation: Group data by UP Master to calculate total videos, average views, and average interaction rate (interaction rate = (likes + coins + favorites)/views) for UP Master ranking and scatter plot analysis.

(4) Interaction Indicator Correlation: Aggregate favorites, likes, and views to analyze their interrelationships through heatmaps.

The "Bilibili Video Analysis System Based on Hadoop+Spark" adopted Spark SQL for time series aggregation and UP Master performance calculation, providing a technical paradigm for the data segmentation link in this project. In particular, its dual-dimensional aggregation strategy of "month + partition" for views effectively supported the implementation of the view trend analysis module [10]. Additionally, Fu et al. mentioned in big data marketing that correlation analysis of user interaction indicators should consider non-linear relationships between indicators. When aggregating favorites, likes,

and views, this project referenced their proposed correlation analysis model, providing a theoretical basis for heatmap display of indicator relationships [11].

5. Visualization Analysis

This project implements two visualization dashboards: the Content and Trend Analysis Dashboard and the UP Master and Interaction Analysis Dashboard, intuitively presenting analysis results through ECharts components.

5.1 Content and Trend Analysis

(1) Tag View Distribution and Popular Tags:

Box plot analysis of tag views reveals that videos under tags such as "migrant worker", "food delivery", and "Bilibili Food Program" exhibit large fluctuations in views with extremely high upper limits. This indicates that these tags gather some hit content, but overall view performance is uneven. It suggests that content creators should not only focus on hot topics when selecting tags but also consider the matching degree between content and tags. Meanwhile, the popular tag word cloud displays the tags of videos with the highest views, intuitively reflecting current hot content areas in the food section. As shown in Figure 1.

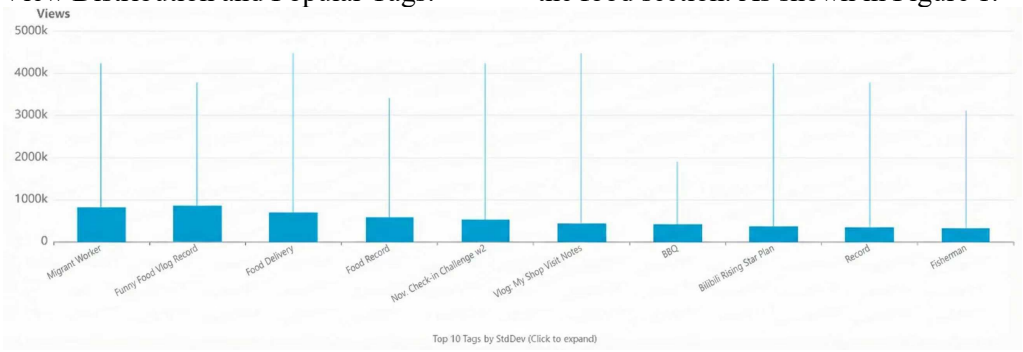

Figure 1. Tag View Analysis Chart

(2) View Time Trend and Release Time Slots: The video view trend chart clearly shows the change in total views of food videos from October 2020 to January 2021. Views gradually increased from October 2020, reached a peak in December 2020, and then declined significantly in January 2021. This trend may be influenced by year-end holidays.

Meanwhile, the time traffic analysis chart shows that videos released at noon have the highest average views and a large number of submissions, making it the most effective release time. It is recommended that content creators optimize their release strategies and prioritize publishing videos at noon or evening. As shown in Figure 2 and Figure 3.

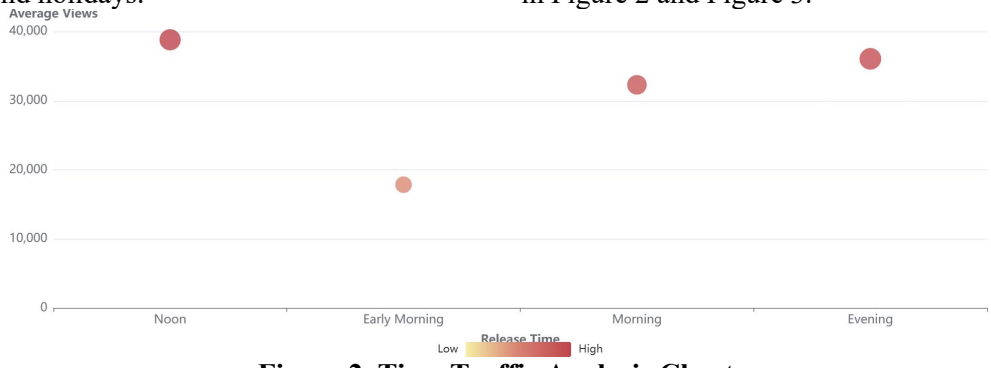

Figure 2. Time Traffic Analysis Chart

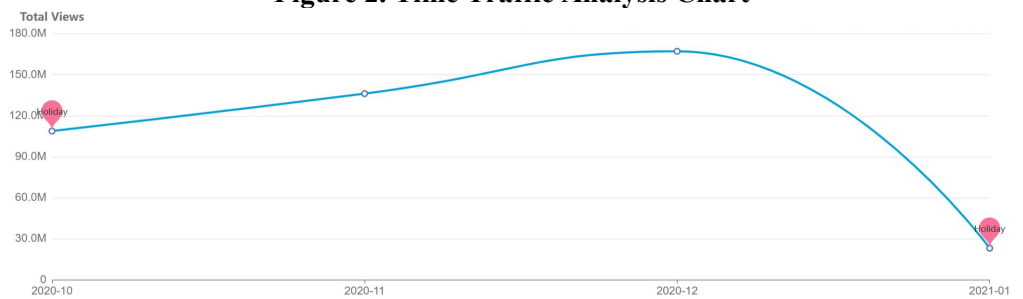

Figure 3. Video View Trend Analysis Chart

(3) Tag Correlation Analysis: The Sankey diagram visualizes the combination and flow relationships between various tags. For example, there is significant content overlap between tags such as "Taste of Life · Annual Food Awards", "rural food", and "production tutorials". This indicates that content producers often adopt a cross-tag combination strategy, which helps expand the audience base and increase content exposure. As shown in Figure 4.

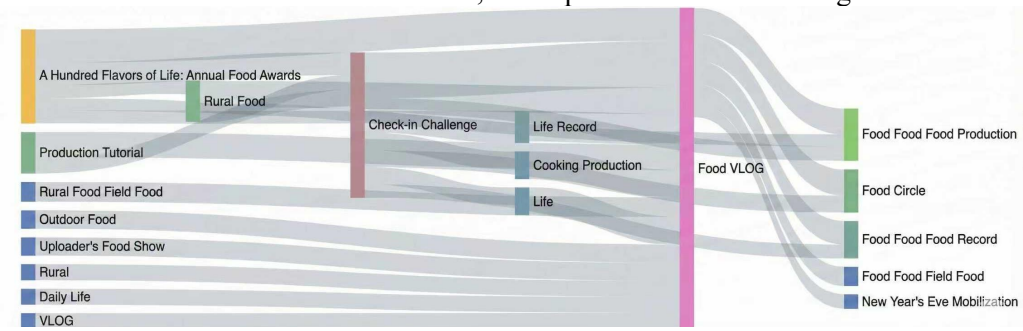

Figure 4. Tag Correlation Analysis Chart

5.2 UP Master and Interaction Analysis

(1) High-Yield vs. Niche UP Master Analysis: The high-yield UP Master ranking chart compares the relationship between the number of videos and average views of UP Masters. Analysis reveals that UP Masters such as "P Niang Jia Dao" (Aunt P Arrives) have far higher average views than others despite publishing very few videos, indicating that their content is more likely to become hits. This suggests that high output does not necessarily lead to high traffic, and content quality and audience matching are more important.

The scatter plot analyzes the average views and average interaction rates of niche UP Masters, revealing that a small number of UP Masters achieve extremely high interaction rates even with low views. Such content exhibits strong community activity and fan loyalty, and the platform may consider further supporting these creators. As shown in Figure 5.

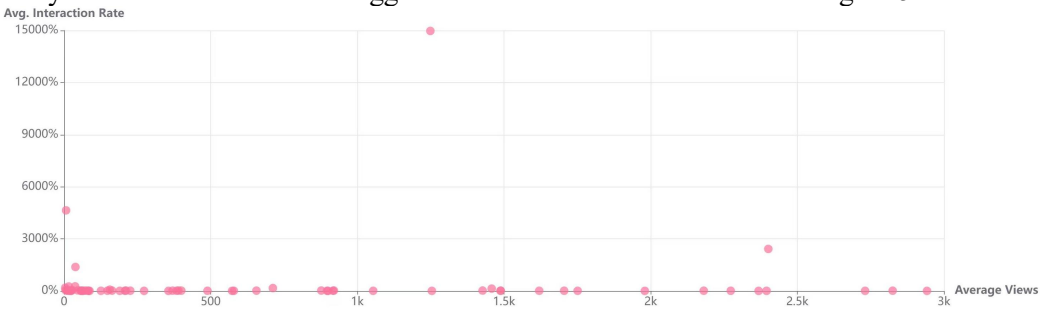

Figure 5. Niche Hidden Gem UP Masters Chart

(2) Video Interaction Data Insights: The radar chart comprehensively displays video performance across multiple core interaction indicators (likes, coins, favorites, interaction rate), enabling effective identification of high-interaction, high-quality content. Additionally, the heatmap, with favorites and views as coordinates, reveals that videos with high favorites generally perform well in views, presenting the characteristic of "high-quality content gathering high interaction and high exposure". This indicates that improving content collection rates is an effective way to drive rapid growth in views. As shown in Figure 6.

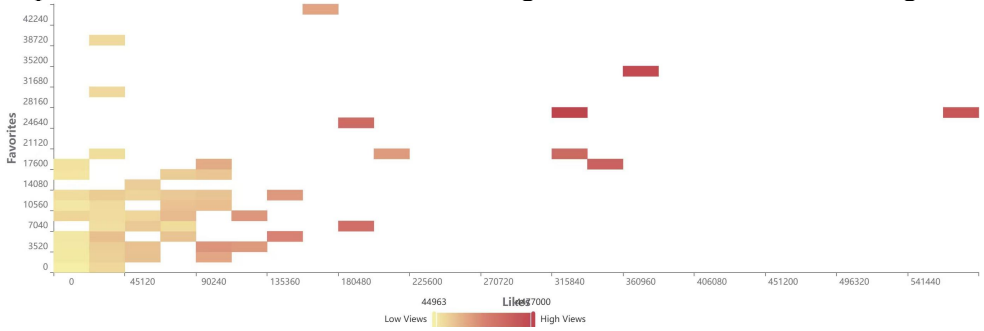

Figure 6. Video Interaction Data Insights Chart

Chen proposed in his research on data visualization design that the combined use of radar charts and heatmaps can effectively display multi-indicator correlation relationships. When analyzing video interaction data, this project drew on his chart combination strategy—using radar charts to identify comprehensively high-quality content and heatmaps to verify the positive correlation between "favorites and views", making the analysis conclusions more convincing [12].

6. Conclusions

Based on big data computing cluster technology and modern front-end and back-end architecture, this project successfully conducted multi-dimensional statistical analysis and visualization of video data in Bilibili's food section, constructing a complete data analysis application. The project adopts a combined architecture of Vue 3, ECharts, Hadoop, Spark, MySQL, and Servlet, ensuring high performance, modularization, and good scalability of the system. Through ECharts charts, in-depth analysis of key indicators in the video ecosystem—such as UP Master influence, content hotspots, and user interaction behaviors—is realized. The system achieves a functional closed loop from data processing to front-end display, integrating practical features such as user registration/login and data export.

However, the project also has limitations: for example, the deduplication logic in the data cleaning link needs further improvement (currently only based on title, publisher, and release time), and the initial strict filter conditions may have led to the loss of valid data. Additionally, the existing system faces issues of long chart component initialization time and decreased rendering performance under large data volumes. Future work will focus on introducing Vuex/Pinia for data caching and optimizing chart lazy loading or on-demand loading mechanisms. Furthermore, more machine learning algorithms may be integrated, such as using clustering algorithms for UP Master segmentation or constructing regression models to predict video hit potential. At the same time, user permission management and automatic/manual data refresh functions should be improved to enhance system practicality and reliability.

References

[1] Feigua Data (Bilibili Edition). Bilibili Food Industry Marketing Insight Report. 2021.

[2] Bilibili. Bilibili Annual Food Data Report. 2022.

[3] Zhou YJ. Research on the Interface Application of Bilibili APP from the Perspective of Emotional Design. Anhui Polytechnic University, 2024. DOI:10.27763/d.cnki.gahgc.2024.000074.

[4] Yu FY, Sun JW, Li RQ. Data Analysis and Development Suggestions for University Libraries Operating Bilibili Accounts. Media Forum, 2024, 7(03):115-117.

[5] Cai W, Zhang JW, Hu GP. Comprehensive Evaluation of Internet Fitness Teaching Videos Based on Big Data Analysis—Taking Bilibili Videos as an Example// Chinese Sports Science Society. Proceedings of the 13th National Congress of Sports Science—Wallpaper Exchange (Sports Statistics Branch). Huaqiao University; Guangzhou Sport University, 2023:262-264. DOI:10.26914/c.cnkihy.2023.081031.

[6] Luo YM. Design and Implementation of a Responsive Data Visualization Platform Based on Vue. Information Technology and Informatization, 2020(09):20-22.

[7] Jing GW, Huang DC. Research on Data Visualization Based on ECharts. West China Broadcasting TV, 2022, 43(20):227-230+234.

[8] Shi YF. Construction of Video Evaluation Indicator System Integrated with Text Data. Zhongnan University of Economics and Law, 2022. DOI:10.27660/d.cnki.gzczu.2022.001585.

[9] Bilibili Popular Video Data Visualization Analysis System Based on Big Data. Juejin, 2025. https://juejin.cn/post/7558233550472118287.

[10] Analysis and Optimization of Bilibili Video Recommendation and Popularity Prediction System Based on Python + Machine Learning. Juejin, 2025. https://juejin.cn/post/7544077353371877427.

[11] Fu XR, Chen J. Big Data Marketing. Posts & Telecom Press, 2023:264.

[12] Chen GQ. Research and Analysis of Data Visualization Design. Electronic Components and Information Technology, 2024, 8(09):151-153+157. DOI:10.19772/j.cnki.2096-4455.2024.9.045.