# Design and Implementation of a Media Knowledge Base Based on User Profiling Analysis (Postprint)

**Authors:** Ma Ming, Chen Xinyi, Chen Jun

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

## Abstract

The development of Internet, big data, and new media technologies has engendered revolutionary transformations in media communication channels and content forms. Analyzing user news adoption and dissemination behaviors across various channels constitutes a crucial component in mining user behavior patterns, constructing user profiles, and enhancing both the dissemination efficacy and adoption rates of editorial content. However, propelled by new media technologies, media convergence has emerged as a prevailing trend, rendering the relationship between users and media increasingly intricate and dynamic. Conventional media statistical service methodologies prove inadequate in satisfying user demands for intuitive, visualized, and multi-dimensional representations of adoption scenarios. This study aggregates information from diverse sources—including users, institutions, and media organizations—and leverages big data analytics, rule-based matching, and tag extraction technologies to construct a media knowledge base grounded in user profile analysis. This initiative fully unleashes the latent potential of data resources as essential factors of production, dismantles information silos across system projects, fosters collaborative data management and service provision, and empowers high-quality data resources to deliver robust support and assurance for numerous R&D products. The paper elaborates on the significance and design principles underlying the media knowledge base, conducts comprehensive investigations into pivotal technologies and their implementation, and consequently proposes prospective intelligent applications built upon this knowledge base.

## Full Text

# Design and Implementation of a Media Knowledge Base Based on User Profile Analysis

**Ma Ming, Chen Xinyi, Chen Jun**
*(Xinhua News Agency Communication Technology Bureau R&D Center, Beijing 100083)*

### Abstract

The development of internet, big data, and new media technologies has brought revolutionary changes to media dissemination channels and content forms. Analyzing how users adopt and disseminate news across different channels constitutes a crucial component for mining user behavior, constructing user profiles, and enhancing both the dissemination power and adoption rates of news articles. However, with the advancement of new media technologies, media convergence has become an inevitable trend, rendering the relationship between users and media increasingly complex and variable. Traditional media statistical services can no longer satisfy users' demands for intuitive, visualized, and multi-dimensional analysis of adoption patterns. This paper constructs a media knowledge base based on user profile analysis by collecting information from multiple sources—including users, institutions, and media organizations—and leveraging big data, rule matching, and tag extraction technologies. This initiative fully taps into the potential of data resources as essential production factors, breaks down information barriers between systems and projects, promotes collaborative data management and services, and enables high-quality data resources to provide robust support for more R&D products. The paper introduces the significance and design principles of the media knowledge base, conducts in-depth research on key technologies and implementation methods, and proposes prospects for intelligent applications based on the media knowledge base.

**Keywords:** Media Knowledge Base; Media Convergence; User Profile; Statistical Adoption

---

According to the China Internet Development Statistics Report released by CNNIC in 2021 [1], by December 2020, China' s internet user base had reached 989 million, with a penetration rate of 70.4%. As the internet market continues to expand, audiences are increasingly fragmented across channels, with declining attention to any single channel, making the user-media relationship more complex [2]. Consequently, to establish user profile models, integrate data resources, and actively respond to the deep integration of media, this paper proposes a model for constructing a media knowledge base based on user profile analysis. Only by incorporating users into the convergence system can we truly build a full-process, holographic, all-staff, and all-effective integrated media platform and secure a competitive position in the media landscape.

### 1.1 Necessity of Constructing a Media Knowledge Base Based on User Profile Analysis

Establishing user profile models provides decision-making references for precise article 推送 and enhanced dissemination impact. With the rapid development of internet and mobile internet technologies, media has entered an era of accelerated convergence. Changes in information distribution channels have altered audience access patterns, making traditional media statistical services inadequate for meeting users' needs for intuitive, visualized, and multi-dimensional analysis of media adoption. Meanwhile, new media technologies have made media convergence a trend, further complicating the user-media relationship. Social media platforms such as Weibo, WeChat, Douyin, and Kuaishou have become channels for news dissemination. To deliver news more efficiently and expand their influence, numerous traditional media organizations have built their own media matrices, forming a "one-point, multi-winged" dissemination pattern.

To fully exploit the potential of data resources, break down information barriers between traditional systems, and better leverage the foundational and innovative roles of data, this paper constructs a knowledge base centered on the media matrix. Simultaneously, to effectively utilize big data technology and core data resources for analyzing user news adoption and dissemination across channels, exploring the depth and breadth of media dissemination, maintaining brand image, and enhancing influence, this paper introduces the significance and design principles of the media knowledge base based on user profile analysis, conducts in-depth research on key technologies and implementation, and proposes prospects for intelligent applications based on the media knowledge base.

### 1.2 Feasibility of Constructing a Media Knowledge Base Based on User Profile Analysis

First, after several years of construction, the statistical monitoring system has accumulated hundreds of millions of article dissemination records, with collected media instances reaching over 330,000. Concurrently, the distribution system has accumulated substantial user subscription data. Despite information barriers between different systems, this vast amount of structured data provides a solid foundation for building the media knowledge base. Second, the overarching trend of media convergence has gradually generated industry expertise. Wang Yijiao [3], using "Huali Daily" as an example, constructed a new media matrix, introduced multi-platform operations, and detailed user feedback across the matrix. Chen Xinglan [4], while presenting media convergence cases, identified three conceptual pitfalls in media matrix construction and emphasized the critical role of users in media convergence. Liu Jing et al. [5] examined the construction of government new media matrices, introducing their composition and development challenges within government systems. These practices provide rich modeling methodologies and application cases for constructing a user profile-based media knowledge base.

## 2.2 Dimensions of the Media Knowledge Base

The media knowledge base aims to construct a media knowledge matrix based on three attributes: user, institution, and media. User data originates from registered users of the Xinhua News Agency Communication Technology Bureau' s distribution system. Institution data is sourced from third-party resources, big data crawling, and manual collation. Media data derives from third-party resources, internet collection, and the media library of the statistical monitoring system. The following sections introduce the data acquisition and attributes for institutions and media.

**Concept 1: Institution** refers to government agencies, organizations, or other work units.

**Concept 2: Media** refers to information dissemination mediums, including traditional and new media, which belong to institutions and possess distinct channel-specific identification attributes.

The core of constructing a user profile-based media knowledge base involves tagging data across user, institution, and media dimensions to build a label system from different perspectives, making the media knowledge base more concrete, reliable, and comprehensive.

The primary indicators of the media knowledge base are designed as three distinct indicator systems corresponding to the three concepts. For institutions, the system primarily examines basic information, regional attributes, and influence attributes. Basic information records authoritative institutional details, regional attributes document geographical distribution, and influence attributes comprise tags such as establishment duration, central/local classification, and commercial/non-commercial status to enable more comprehensive institutional analysis. For media, the system focuses on basic information, regional attributes, and channel attributes. Channel attributes establish unified media identifiers and extract channel-specific information based on different platforms including websites, "two micros and one terminal," overseas social platforms, short-video platforms, long-video platforms, and public accounts. For users, the system examines basic information and subscription details. Table 1 illustrates the main indicator information.

Building the media knowledge base entails constructing a media matrix to establish correlations among users, institutions, and media, discovering relationships between institutions and media, and completing information on subordinate media and their channels to support subsequent user profiling and personalized recommendations.

The essence of constructing a user profile-based media knowledge base is to fully utilize user-institution-media data, visualize user needs according to the media matrix, and apply this to statistical adoption services to assist editorial decision-making and achieve precision services. The user profile-based media knowledge base model comprises three layers: data layer, analysis layer, and application

layer, as shown in Figure 3 [Figure 3: see original paper].

Media data is categorized into traditional and new media. Traditional media such as newspapers, television, radio, and websites still account for a significant proportion of news transmission and adoption. Meanwhile, as users' information consumption patterns evolve, "two micros and one terminal," Douyin, Kuaishou, and other short-video platforms, as well as Bilibili, Tencent Video, and other long-video platforms, have gradually become the main battleground for public opinion.

This paper has compiled mainstream and influential media channels from domestic and overseas social media to provide sample support for subsequent internet media information collection and the introduction of third-party TRS resources. The main media channels are shown in Figure 2 [Figure 2: see original paper].

**Figure 2** Main Media Channels

The data layer serves as the foundation for building the media knowledge base, comprising data sources and data collection. Data sources are primarily obtained from the distribution system, statistical monitoring system, and third-party resources. Data collection gathers adoption data from websites, WeChat, Weibo, clients, and other third-party media through internet resource introduction projects and county-level media convergence collection teams. The collected media data is serialized and stored in a raw database, which is regularly updated and refined based on media 入库 status to establish user-institution-media relationships. Finally, through data cleaning, transformation, reduction, and integration, the system prepares data for further analysis in the analysis layer.

The analysis layer builds the media knowledge base' s label system by combining static data (region, qualifications) of institution-media relationships with dynamic data (distribution frequency, comments, reads, adoptions) using behavior analysis, clustering analysis, association analysis, and tag extraction technologies. Through label modeling and analysis, the system can uncover individual and group user characteristics to support personalized intelligent services in the application layer.

The application layer proposes personalized services based on the media knowledge base, aiming to provide decision-making references for editors through big data analysis, optimize column structures, deliver precise and personalized recommendations to users, and assist marketers in copyright protection and potential user discovery.

## 2.3 Implementation of the Media Knowledge Base

### 2.3.1 Establishing Data Synchronization Mechanisms and Storing Structured Data

Beyond introducing third-party resources, the media knowledge base has established data synchronization mechanisms with the distribution and statistical monitoring systems. User subscription statuses and 入库 media data are updated daily to timely 完善 user-institution-media relationships and ensure the timeliness of the media knowledge base.

### 2.3.2 Big Data-Driven Matching to Reduce Manual Processing

Through preliminary data cleaning, clustering, and analysis of 5.9 million entity data records, we obtained over 87,000 institution names related to news (19,701 institutional names and 68,153 company names). However, the 87,000+ institution names cannot be used directly. Since the 5.9 million data records were entity words obtained through big data-based segmentation algorithms, entity identification depends on the accuracy of these algorithms. Nevertheless, the 87,000+ institution names remain highly valuable. On one hand, the results cover all aspects of the media industry, providing a relatively comprehensive list of domestic institutions and companies. On the other hand, institution and company naming conventions are relatively standardized, with comprehensive variations of the same entity' s name. Therefore, we establish user-institution relationships primarily through programmatic matching rules, supplemented by manual adjustments.

The keyword clustering result distribution for news-related institutions is shown in Figure 1 [Figure 1: see original paper].

**Figure 1** Keyword Clustering Result Distribution

### 2.3.3 Providing Association Query Services to Meet User Profiling Needs

Constructing a user profile-based media knowledge base involves fully leveraging big data and rule matching technologies to establish a comprehensive and 完善 user-institution-media label system. After absorbing data from various sources including internet resource introduction, statistical monitoring systems, distribution systems, and third-party resources, the media knowledge base operates as an independent system providing external query services. These services include media information queries, institution information queries, media matrix queries, and user-institution-media adoption queries, providing data support for subsequent intelligent applications such as user profiling, copyright detection, and article recommendation.

## 3. Media Knowledge Base Application: A Case Study of County-Level Media Convergence Center Users

To accelerate county-level media convergence center integration and actively respond to central government requirements for strengthening county-level media convergence construction, this paper collected and established media matrices for 328 county-level media convergence centers nationwide, generating a total of 1,241 media outlets—an average of four dissemination channels per center. By combining data from county-level media convergence center users, the media knowledge base, and article adoption statistics from 2020, we developed preliminary user analysis for county-level media convergence center users.

### 3.1 User Profiling

Building user profile models involves using media subscription information and adoption records to create tags representing user preferences and habits. The collection of tags from numerous users forms user groups with shared characteristics, providing effective support for targeted media preference activities in integrated media knowledge bases. The following analysis examines county-level media convergence center users across four dimensions:

First, the dissemination channels of county-level media convergence center users are primarily focused on "two micros and one terminal," with fewer short-video platforms like Douyin and Kuaishou, and insufficient account operations on long-video platforms such as Bilibili and Tencent Video.

Second, analyzing article adoption patterns reveals that text and image-based articles remain dominant due to their high information density and low storage requirements, while video article adoption remains limited. This indicates that editors need to strengthen media convergence capabilities by combining text, audio, and video production, while convergence centers must enhance platform operation capabilities to increase news dissemination forms and attract audiences.

Third, in column ranking details, the "Urban-Rural Development" column achieves the highest adoption rate at over 80%, followed by "General Secretary Reports" and "Current Affairs News," both exceeding 70%. Lower-ranked columns include "Sports Express," "Health Encyclopedia," and "Financial Focus." This demonstrates that county-level media convergence center users prefer local current affairs news over sports, finance, and other themes.

Fourth, regional attributes of county-level media convergence centers show that among the top 50 users by adoption volume, county-level media convergence centers from Inner Mongolia, Gansu, and Beijing rank higher. This indicates that in these three regions, county-level media convergence centers developed earlier and established relatively 完善 media convergence dissemination mechanisms, and that Xinhua News Agency' s promotion and influence are comparatively greater in these areas.

### 3.2 Personalized Recommendation

By constructing user profiles, discovering user behavior patterns, and understanding user content preferences, we can provide personalized recommendations for different user groups. Based on the aforementioned column rankings and regional attributes, recommendations prioritize the "Urban-Rural Development" column, county-level media convergence center-specific lines, and local articles, providing effective technical support for enhancing user stickiness.

### 3.3 Copyright Monitoring

The media knowledge base provides user adoption, institutional adoption, and media adoption data to the copyright monitoring system. Through technical means such as big data extraction, cloud recognition, and information comparison, the system monitors copyrights of articles or digital information, helping to promptly identify and investigate infringement and piracy, discover infringement clues, and obtain compelling electronic evidence for effective copyright protection. Additionally, through the aforementioned user profiling and personalized recommendations, the system can provide users with creative materials and preferred articles, promoting news dissemination and enhancing the agency' s influence.

## 4. Conclusion

The media knowledge base has established relationships between users and media, collected and aggregated global institutional data and their various channel subordinate media, and constructed a hierarchical, multi-channel media matrix knowledge base. The system's highlights include the expansion and improvement of media channels, the establishment of data-driven thinking, the breakdown of information silos between systems, and the implementation of efficient and rapid information query methods, providing data support for subsequent intelligent applications based on the media knowledge base.

## References

[1] CNNIC. China Internet Development Statistics Report [R]. Beijing: China Internet Network Information Center, 2018.

[2] Dai Jianyun. Building Media Matrix Dissemination Power in the Era of Media Convergence [J]. China Broadband, 2021(7): 191.

[3] Wang Yijiao. Preliminary Exploration of Private Domain Traffic Operation in the Publishing Industry—Taking the "Huali Japanese" New Media Matrix as an Example [J]. Modern Publishing, 2021(2): 85-88.

[4] Chen Xinglan. Three Conceptual Pitfalls in "Media Matrix" Construction [J]. Media, 2020(11): 65-67.

[5] Liu Jing, Ling Yimin. Analysis of Government New Media Matrix Construction in China [J]. Publishing Wide Angle, 2020(19): 23-25.

[6] https://github.com/nirenxiaoxiao/Company-Names-Corpus

**Authors:** Ma Ming (1993-), female, from Handan, Hebei, Master' s degree, Engineer, research direction: algorithm applications; Chen Xinyi (1986-), female, from Fujian, Master' s degree, Engineer, research direction: big data in journalism; Chen Jun (1977-), female, from Sichuan, Master' s degree, Senior Engineer, research direction: data analysis.

*(Editor: Zhang Xiaojing)*

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