Analysis of Post-Prints Based on Big Data Applications in Contemporary Publishing

Authors: Ding Yanwei, Wu Xiayan

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

Abstract

With the development of science and technology, the modes of information dissemination, information reception, and media have undergone tremendous changes. Due to the increasingly widespread application of new media technology in media, people' s daily lives are becoming increasingly inseparable from it. For the publishing industry, big data runs through all aspects of the publishing process, from topic planning, production of in-house resources, layout production to marketing promotion. Therefore, media organizations must expedite digital transformation, leverage big data in accordance with their specific circumstances, and utilize it to drive business process transformation and business model innovation. Based on this, this paper briefly analyzes the current application of big data in publishing, with the aim of providing recommendations for the development of the publishing industry.

Full Text

Analysis of Big Data Applications in Contemporary Publishing

Ding Yanwei, Wu Xiayan
(Capital Normal University Press, Beijing 100048)

Abstract: With the advancement of science and technology, the modes of information dissemination, reception, and media have undergone profound transformations. As new media technologies become increasingly pervasive, people' s daily lives have grown inseparable from these innovations. For the publishing industry, big data permeates every stage of the publishing process, from topic planning and content production to layout design and marketing promotion. Consequently, media organizations must accelerate digital transformation, integrate big data with their actual operations, and leverage it to drive business process reforms and business model innovation. This paper provides a brief

analysis of current big data applications in publishing, aiming to offer recommendations for the industry' s development.

Keywords: big data; publishing; film and television media; new media technology; application analysis

Classification Code: G250.7
Document Code: A
Article ID: 1671-0134(2022)03-074-03
DOI: 10.19483/j.cnki.11-4653/n.2022.03.023
Citation Format: Ding Yanwei, Wu Xiayan. Analysis of Big Data Applications in Contemporary Publishing [J]. China Media Technology, 2022(03): 74-76.

1. Overview of Big Data and Big Data Technology

Big Data, once a keyword in the internet information technology sector, has now entered mainstream discourse. Against the backdrop of technological revolution and industrial transformation, big data—whether as a strategy, tool, or resource—possesses tremendous transformative power that impacts numerous industries and fields. The publishing industry, as a vital component of the cultural and information sectors, is no exception. Big data has created a complete cycle throughout the publishing process, covering everything from topic planning and internal product production to production design and marketing. It has not only penetrated every functional domain of publishing but has also become a crucial guarantee for publishers to enhance productivity, innovation, and competitiveness.

Historically, there was no unified definition of Big Data. It was generally defined as massive data, voluminous data, or enormous information resources [1], referring to datasets whose scale and volume are too large for humans to intercept, manage, process, and organize into readable information. Big data has four main characteristics: large volume, diverse data types, fast processing speed, and low value density. In the big data era, the key question is how to mine and utilize such vast amounts of data, which directly influences the application of big data technologies. These applications can be categorized into big data mining, processing, storage and management, analysis, application, and security, with big data analytics being the most critical. Precise data analysis forms the foundation of data application and ensures that big data realizes its value and delivers benefits to users.

From this perspective, key big data technologies include cloud computing, distributed file systems, and parallel computing architectures. Big data originates from and is closely related to cloud computing: big data provides analytical content for cloud computing, while cloud computing offers the infrastructure for big data analysis. As data volumes escalate to PB (1024TB = 1PB), EB (1024PB

= 1EB), or even ZB (1024EB) levels, large amounts of redundant or invalid data inevitably emerge, requiring substantial computing power and scalable collection capabilities—precisely what cloud technology provides [2]. Distributed file systems can connect physical resources to remote nodes through computer networks, enabling multiple nodes to form a file system network. Parallel computing allows simultaneous multiple computations. Both distributed file systems and parallel computing databases represent key technologies in the big data era, with the Hadoop architecture being a typical implementation model.

2. Development Background of Big Data Applications

Both domestic and international media referred to 2013 as the "Year of Big Data,"when it became a hot topic. Prior to this, numerous research achievements related to big data had already emerged, with the most influential being a book, a research report, and a government development plan that appeared across three timelines spanning over 30 years.

First, in March 1980, American futurist Alvin Toffler published *The Third Wave* [3], in which he described his predictions and visions for the future, first citing big data and regarding it as the "colorful movement of the third wave." Second, in 2001, the concept of "big data" first appeared in a Gartner research report, which defined its characteristics as high volume, high variety, and high velocity [4]. Third, in 2012, the Obama administration launched a "Big Data R&D Initiative" involving six federal agencies with an investment of $200 million, elevating big data development to the national strategic level for the first time [5].

In China, the development environment for big data applications has been favorable since 2013, with progress roughly divisible into three stages:

Stage 1: 2013—Exploration and Proof-of-Concept Phase
In 2013, some studies provided theoretical explanations for the importance and implementation of large-scale data processing technologies. For example, in July 2013, the 2012-2013 China Digital Publications Annual Report was released at the Fifth China Digital Publications Exhibition. This report conducted extensive data analysis and mining on digital publications, identifying it as one of the development trends in digital publishing.

Stage 2: 2014—True Implementation Phase
In 2014, China introduced big data policies and measures, with support from government departments such as the Ministry of Industry and Information Technology, the National Development and Reform Commission, the Ministry of Science and Technology, and the Ministry of Finance in terms of funding and development planning. Governments at all levels also played important roles in large-scale information projects, research initiatives, and technology dissemination. The Zhongguancun Big Data Exchange Industry Alliance was established, and big databases began to be constructed in provinces and cities nationwide. The release of *Opinions on Accelerating the Cultivation of Big Data Industry*

Clusters to Promote Industrial Transformation and Upgrading provided standards for the management and control of large-scale data transactions in China.

Stage 3: 2015 to Present—National Strategic Planning Phase
At the national strategic development level, big data promotion began in 2015. The State Council issued the *Outline for Promoting the Development of Big Data* (hereinafter referred to as the *Outline*) [6]. The *Outline* positioned data as a major national strategic resource, making the accelerated introduction and deepening of big data a necessary requirement and inevitable choice.

3. Big Data Applications in the Publishing Process

The most common description of the traditional publishing process is "author—publisher—reader," where the author is the subject, the reader is the terminal, and the publisher is the intermediary. Copyright and orders form the absolute center, with readers merely serving as content recipients. The greatest drawback of this open linear structure is the separation of demand and production into two distinct entities, where experience becomes the primary basis and empirical connections lack necessary evidentiary support.

The largest application of big data in publishing is creating a "reader—publisher—author—reader" closed-loop structure. Readers serve as both the starting point and endpoint, participating as both content producers and consumers. This represents a typical interpretation of audience-centered theory in the new media era and embodies the publishing philosophy under big data thinking. This concept and practice originated in e-commerce, most notably represented by Amazon internationally and JD.com in China. By collecting and analyzing user behavior data (including searches, views, purchases, votes, etc.), the direct connection between production and demand ensures profit maximization.

3.2 Topic Planning Based on Big Data

Big data first facilitates topic selection and planning, which is particularly important for new media companies involved in publishing. These companies typically possess robust online data and have advantages in collecting, analyzing, and utilizing structured, semi-structured, and unstructured data related to traditional publications. User behavior can be recorded at any time through social platforms and e-commerce platforms, quickly and accurately reflecting social hotspots and trends in specific fields to provide important recommendations for topic planning. In recent years, the success of many bestsellers has been based on incorporating large-scale data into topic planning. The most representative example in China is *David Beckham*, launched by JD.com in 2014 based on analysis of 17 million website users [7].

3.3 Mass Production Based on Big Data

During content creation, big data serves as the optimal framework for determining how authors should express their work' s content or historical development direction. Electronic publishing company Coliloquy has achieved success in this area. Coliloquy uses Amazon' s Kindle data developer program to collect user data, particularly content that users repeatedly highlight and analyze. By analyzing and extracting topics, Coliloquy determined its publishing direction toward youth, romance, and science fiction, openly recruiting writers to join its team and recently adding crime and legal thrillers. The "Coliloquy model," which is at the center of the hottest discussions, is essentially thematic publishing based on data analysis of published works. Ninety percent of readers have read these books (typically priced between $2.99 and $7.99), and 67% have reread them. The success of the "Coliloquy model" is based on thematic positioning through analysis of published data [8].

3.4 Presentation and Production Based on Big Data

Layout production mainly includes content review, editing, correction, and typesetting. In recent years, digital production platforms based on XML data processing standards have been applied in internet environments. These include multi-user online editing platforms for multi-purpose remote editing collaboration, providing opportunities for real-time communication and editing applications among various roles including authors, readers, editors, and publishers. Additionally, finished product data and fragmented data from the production process can be stored simultaneously, facilitating content tracking and extraction. During the editing process, digital annotation tools can electronically mark manuscripts according to commonly used predefined correction symbols. Massive databases ensure that content can be matched with custom publishing styles, automatically typeset, linked to different templates, and created in different versions. Editing based on big data not only improves editing efficiency but also enhances the quality of final products.

3.5 Precision Marketing Based on Big Data

Precision marketing aims to "reduce marketing costs and improve marketing effectiveness" by delivering products to users who truly need them. By using extensive information technology for marketing, publishers and media can not only deepen customer data but also utilize various platforms such as social networks to maintain personal and interactive connections, thereby increasing or improving user loyalty. This involves analyzing social network user circles and implementing targeted marketing campaigns.

Amazon excels at marketing data. The company replaced its previous expert recommendation system with a personalized data-driven recommendation system, thereby boosting sales. The system recommends books to readers by analyzing consumption information (such as book purchases and book follows). In

addition to this customized recommendation system, Amazon has also entered the important stage of marketing and data delivery. Its physical pre-order distribution utilizes big data technology to conduct in-depth analysis of past consumption expenditures, search history lists, and predictions of new product purchases. Products are prepared for delivery to customers or stored near customers in advance, so that when customers place orders, they receive goods in hours rather than days. This silent delivery model can partially improve customer loyalty and enhance Amazon' s reputation among customers.

Overall, big data applications in the publishing industry remain in their infancy, with many issues requiring exploration and verification. The most important task for domestic media enterprises is to accelerate digital transformation, develop big data applications based on their actual conditions, and utilize big data to drive business process transformation and business model innovation.

4. Big Data Applications in Current Publishing Models

4.1 Thematic Publishing Based on Data Analysis

The use of big data and analytical methods aims to discover principles, identify patterns, and predict applications through data analysis, with thematic publishing being one such application. The Google Books database has collected relevant publication data from the Common Era to the 20th century. By analyzing data from various disciplines, particularly extracting and classifying high-frequency topics, it holds significant commercial value. The "Coliloquy model," currently at the center of the hottest discussions, is also essentially thematic publishing.

4.2 Visual Publishing Based on Data Interaction

Visualization technology was first applied in computer science, using computer graphics and image processing technology to convert data into graphics, audio, video, or animation for interaction with machines. It is a technology that integrates data presentation, processing, and decision analysis. Reading experience is an important indicator for evaluating current publications. Visualization based on big data interaction can not only present various abstract and complex knowledge more intuitively and simply, largely eliminating reading barriers and providing efficient and convenient reading, but also enable simultaneous data display across multiple spaces, bringing people a 3D reading experience. Visual publishing represents the latest model in the publishing industry and will have a revolutionary impact on its development. Currently, this model is suitable for children' s publications and technology publications, with the publishing direction combining flat and stereoscopic, static and dynamic elements.

4.3 Interactive Data-Driven Publishing

One of the greatest achievements of the Web 2.0 era was the birth of Wikipedia, which represents the most successful application of wiki technology in practice. On the same open data platform, users interpret the same events or viewpoints from different angles, and the demand for personal interpretation is a key characteristic of new media audiences. This concept and practice has given rise to a new publishing model—interactive publishing. Storybird' s "Digital History" creation service platform is an excellent example. Storybird is a public platform based on visual storytelling that provides free illustrations from around the world, encouraging readers to select interesting images and share them in written form, thereby providing extensive services for original books and publications. Personal interpretation of illustrations is a continuation of content interaction, and the fusion of reader inspiration and experience makes each version extremely unique with its own attribution. Storybird has over 2 million users worldwide and has created 5 million stories within two years of its establishment. Through online or offline publishing, it offers a new way of thinking and a new publishing method.

4.4 Collaborative Publishing Based on Data Sharing

In the transformation and development of traditional media, digitalization is the direction and strategic cooperation is the method. Creating digital content and professional databases (especially massive data) is an inevitable trend and requirement for cross-community, cross-social, cross-industry, and cross-boundary cooperation. The "Agricultural Digital Library" adopts an open public model based on shared data. This library is a publishing project jointly developed by multiple provinces. At the Nanchang meeting of the 9-province joint platform, agreements were reached on electronic copyrights for over 1,600 agricultural books, with unified authorization given to Zhongyuan Farmers Press and Jiangsu Phoenix Agricultural Publishing Center. A resource library of 300 agricultural books was supported in the form of open ports, and the provinces jointly built the *Agricultural Digital Library*. The first project compiled 2,000 agricultural books, pooling central resources from nine provinces, while the second project will expand to national agricultural publishing institutions. This data sharing and horizontal publishing model based on data exchange reduces repetitive work in reference linking and allows for content regenerability, diversity, and resource integrity, thereby improving data publishing quality.

In 2013, Alibaba restructured 25 business departments to acquire data from related industries and sectors, enriching its powerful database. In 2014, at a major information conference in Beijing, the founder of Alibaba Group announced in a speech that humanity was transitioning from the IT era to the DT (Data Technology) era. Alibaba has captured substantial data dividends, using data to obtain benefits—a key factor for the future. At the 2015 Hangzhou Yunqi Conference, it was announced that the DT era is a new energy era where "the main source of this period is not oil, but data." By the end of the 2016 Yunqi

Conference, Jack Ma reiterated that "future trends will be driven not only by knowledge but also by intelligence and data." "The future based on the internet and massive data technology creates countless imaginations and spaces for humanity." Although most publishing companies do not have abundant information resources, advanced technology, or sufficient resources, they should embrace the concept of using big data to control publishing development. Of course, big data technology applications in the publishing industry remain in their infancy, with many issues requiring exploration and verification. However, its value and importance provide unlimited imaginative space for the publishing industry' s development.

References

[1] Sun Ying. Analysis of Development Trends in the Digital Publishing Industry in the Big Data Era[J]. Communication Power Research, 2019(36): 210.
[2] Cai Yunxuan. Analysis of Innovation in Power Marketing Management Based on Big Data[J]. Economic and Trade Practice, 2019(12): 109.
[3] Xie Hongming, Yang Hao. Research Trends and Frontiers of Big Data in Business—A Bibliometric Analysis Based on Scientific Knowledge Mapping[J]. Science and Economy, 2019(5): 42-46.
[4] Liu Xinglei. Discussion and Application Analysis of Radio and Television Network Transformation Based on Big Data[J]. China New Communications, 2019(18): 101-101.
[5] Yang Xi. Publishing Culture and Editorial Role Transformation in the Big Data Era[J]. China Media Technology, 2021(1): 3.
[6] Yan Chengzhen, Song Di. "Big Data" Era May Trigger a New Pattern of Media Development[J]. China Media Technology, 2012(10): 2.
[7] Li Li. On the Application of Big Data in Publishing Social Media[J]. News Research Guide, 2019(10): 193-194.
[8] Li Dongyan. Analysis of the Application of Big Data in Library Management and Services[J]. Science and Technology Economic Guide, 2019(4): 168.
[9] Hao Yang. Implementing the National Big Data Development Action Plan and Serving Digital Publishing Transformation and Upgrading with Big Data Thinking[J]. Science-Technology & Publication, 2016(1): 23-24.

Author Biographies:
Wu Xiayan (1989-), female, from Gaomi, Shandong, holds a master' s degree and works as an editor. Her research focuses on education and linguistics.
Ding Yanwei (1970-), female, from Shijiazhuang, Hebei, is an on-the-job graduate student and associate editor. Her research focuses on primary and secondary school Chinese teaching materials and methods, as well as editing and publishing.

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