Journal of Information Systems Engineering and Management
2025, 10(61s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Big Data-Based Recommendation Systems

Luís Barata[1][2], Diogo Mata[1], Jorge Nunes[1], Lucas Rodrigues[1]
[1]*Escola Superior de Tecnologia - Instituto Politécnico de Castelo Branco, Castelo Branco, Portugal*
[2]*Instituto de Telecomunicações, Universidade da Beira Interior, Covilhã, Portugal*

ARTICLE INFO

Received: 26 Dec 2024

Revised: 14 Feb 2025

Accepted: 22 Feb 2025

ABSTRACT

Introduction: With the exponential growth of data, digital platforms increasingly rely on Big Data technologies to personalize user experiences and improve the accuracy of item recommendations. Recommendation systems play a critical role in e-commerce, entertainment, and social media by analyzing user interactions, behaviors, and preferences. However, the complexity of large-scale data processing and the diversity of filtering techniques pose significant challenges to achieving high performance and scalability.

Objectives: This study aims to analyze how Big Data technologies are applied in modern recommendation systems, emphasizing their role in enhancing personalization and system performance. The work seeks to identify the main data collection strategies, algorithms, and tools adopted in recent research, as well as to assess how these systems address challenges related to scalability and real-time processing.

Methods: A systematic literature review was conducted using the IEEE Xplore database, focusing on articles published between 2015 and 2025. The search targeted studies combining Big Data and recommendation systems within e-commerce contexts. Out of 49 retrieved publications, 10 met the inclusion criteria, and 5 were ultimately selected after applying exclusion filters. Each selected paper was analyzed regarding its objectives, employed algorithms, data sources, and achieved outcomes.

Results: The reviewed studies demonstrate a wide variety of approaches and technologies, including Hadoop and Spark frameworks for large-scale data processing, deep learning models such as DeepLearning4j for real-time prediction, and classical data mining algorithms like K-Means, Apriori, and FP-Growth. Hybrid methods combining collaborative and content-based filtering were shown to overcome limitations such as the cold start and data sparsity problems. Scalability was addressed through distributed processing and optimization techniques like network pruning and parallel computation. These systems achieved higher recommendation precision and responsiveness across different e-commerce and media applications.

Conclusions: The analysis confirms that Big Data–driven recommendation systems are essential for enhancing user engagement and conversion in digital platforms. By integrating data mining, machine learning, and distributed processing technologies, these systems deliver efficient, adaptive, and context-aware recommendations. Future developments should continue exploring hybrid and deep learning–based approaches to further improve scalability, personalization, and computational performance in increasingly complex digital ecosystems..

Keywords: Big Data, Recommendation Systems, Filtering, E-commerce, Deep Learning.

INTRODUCTION

As the volume of data continues to grow exponentially, digital platforms increasingly leverage this information to enhance user experience and optimize item recommendations. However, the complexity of data and the diversity of recommendation techniques pose significant challenges for these systems. According to (Abbaschian, 2017) and (Kyung-Yong Jung, 2004) recommendation systems can be classified into three main types: *collaborative filtering*, *content-based filtering*, and *hybrid filtering*.

In collaborative filtering, users with similar interests are assumed to interact with items in comparable ways. This method relies on user ratings and interaction histories to identify preference patterns, allowing items to be recommended among users with similar profiles. Nevertheless, this approach faces several challenges, such as the *cold start problem*—which occurs when insufficient data are available for new users or items—and *rating dispersion*, which arises when users with similar profiles assign very different ratings to the same item.

In contrast, content-based filtering focuses on analyzing the characteristics of each item to build profiles for both items and users. As noted by (Abbaschian, 2017) this approach enables the recommendation of new items based on the similarity of attributes to previously rated items, independent of other users' interactions. However, this technique can lead to *over-specialization*, where recommendations become overly restricted to similar items, thus limiting the discovery of diverse content.

Beyond these classical methods, **hybrid filtering** combines the advantages of both approaches. As demonstrated by (Kyung-Yong Jung, 2004) and supported by (Abbaschian, 2017), hybrid systems can mitigate the weaknesses inherent in each individual method. When few ratings are available or when evaluations are highly dispersed, the content-based component provides complementary information through item attributes, reducing the cold start and sparsity issues. Conversely, when over-specialization occurs, the collaborative component allows for greater flexibility, enabling the recommendation of items that are not identical but have been appreciated by users with similar preferences.

In addition to these traditional approaches, advanced *deep learning* techniques have been increasingly incorporated to refine recommendation algorithms. As illustrated in (Personalized Recommendations: How Netflix and Amazon Use Deep Learning to Enhance User Experience, 2025), platforms like Netflix and Amazon employ complex neural networks capable of capturing subtle behavioral patterns and implicit user preferences, leading to highly contextualized and effective recommendations. The systematic analysis presented in this study aims to explore the use of Big Data in recommendation systems, with a primary focus on applications in the **e-commerce** domain. These systems play a key role in content and service personalization and are widely used in e-commerce platforms, social networks, and streaming services, among others. Big Data–based recommendation systems have been shown to positively influence the conversion of visitors into customers by leveraging browsing behavior data to recommend products and reduce *search abandonment* (Johnson, 2024).

METHODOLOGY

To deepen the understanding of Big Data–driven recommendation systems, this study conducted an analysis of scientific publications aimed at addressing a set of key research questions that support a technical discussion of the topic. The following questions were formulated to explore how these systems are developed and how they perform under modern technological demands:

Q1. How are the data used by the system collected?

Q2. What are the main tools and technologies used in recommendation systems?

Q3. How do these systems handle challenges related to scalability and performance?

(i) Research Strategy

To ensure a comprehensive understanding of recent developments in Big Data–based recommendation systems, a literature search was conducted in the IEEE Xplore database (**IEEE, 2025**). The search focused on articles published between 2015 and 2025 — a critical period for the evolution of Big Data technologies and their application in recommendation systems.

The search strategy was designed to capture studies specifically addressing the use of Big Data in recommendation systems within the e-commerce context. After several refinements to align results with the research objectives, the final search string was defined as follows:

("big data" OR bigdata) AND ("recommender system" OR "recommendation system" OR "personalized recommendation") AND ("case study" OR real-world OR industry) AND "e-commerce"

The search was conducted on **April 1, 2025**, yielding a total of **49 results**.

(ii) Inclusion Criteria

The 49 retrieved articles were screened to identify those providing relevant information on the topic under study. The initial selection was based on the analysis of titles and abstracts. Articles were included if they met the following criteria:

1. Explicit mention of one or more recommendation systems, algorithms, or technologies; and
2. Clear reference to the use of Big Data.

After applying these inclusion criteria, a total of **10 articles** were retained for further review.

(iii) Exclusion Criteria

The 10 preselected articles were then read in full to evaluate their objectives, methodologies, and reported outcomes. In this stage, **five articles were excluded** because they lacked sufficient detail regarding the algorithms implemented in their recommendation systems. Consequently, **five scientific papers** were included in the final analysis of this study.

RESULTS

Figure 1 presents a summary of the article selection process. A total of 49 documents were retrieved from the IEEE Xplore database (IEEE, 2025). After applying the inclusion criteria, 39 documents were excluded, leaving a total of 10. Subsequently, based on the exclusion criteria, five additional documents were removed: three because they contained only brief descriptions of the algorithms used, and two because they were overly conceptual.

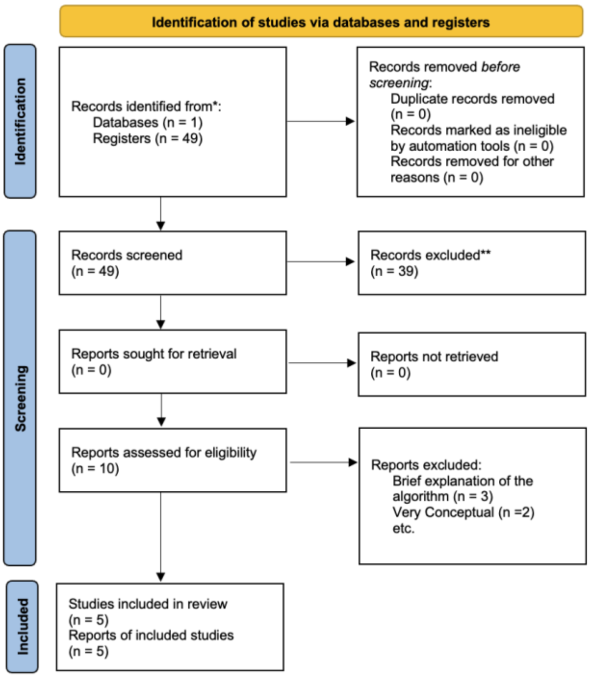
Identification of studies via databases and registers
Identification
Records identified from*:
Databases (n = 1)
Registers (n = 49)
Records removed *before screening*:
Duplicate records removed (n = 0)
Records marked as ineligible by automation tools (n = 0)
Records removed for other reasons (n = 0)
Screening
Records screened (n = 49)
Records excluded** (n = 39)
Reports sought for retrieval (n = 0)
Reports not retrieved (n = 0)
Reports assessed for eligibility (n = 10)
Reports excluded:
Brief explanation of the algorithm (n = 3)
Very Conceptual (n =2)
etc.
Included
Studies included in review (n = 5)
Reports of included studies (n = 5)

Figure 1. Article selection process (adapted from the PRISMA 2020 flow diagram). (Page MJ, 2021)

(i) Characteristics of the Included Studies

The five studies selected for analysis were published between 2015 and 2025. All of them were conducted in China and include a detailed presentation of the algorithms used in their respective recommendation systems.

(ii) Analysis of the Included Studies

Table 1 summarizes the key aspects of each study. For every analyzed work, the table provides an overview of the algorithms employed, the data sources used, the application domain, and the specific objectives of each recommendation system.

Table 1. Summary of the analyzed studies.

Study	Algorithm(s)	Data Source(s)	Application Area	Objective
Mengyi Zhang (2016)	Data Mining, Hadoop Framework	STB, Baidu, Sohu, Sina Weibo, WeChat	Television Broadcasting	Increase audience engagement and optimize TV program recommendations.
Yequn Huang (2024)	Fuzzy Logic, Neural Networks, Network Pruning	User behavior data (browsing, keywords, purchases, clicks)	E-commerce	Improve personalization and optimize marketing strategies.
Shufang Zhao (2023)	Collaborative Filtering, KNN, ALS (Spark MLlib)	User purchase and interaction history	Mobile E-commerce	Enhance recommendation accuracy and reduce data sparsity.
Bai Li (2021)	Deep Learning, MapReduce, Flume, Kafka, Spark Streaming	Real-time user interactions (clicks, navigation, purchases)	E-commerce Platforms	Generate real-time and adaptive recommendations.
Hulin Jin (2024)	K-Means, Apriori, FP-Growth, Collaborative Filtering	Browsing and social media activity logs	Digital Platforms	Improve filtering efficiency and user experience through hybrid recommendation.

The objective of this review is to understand how the various recommendation systems described in the selected studies operate.

The article **"A TV Program Recommendation System Based on Big Data"** (Mengyi Zhang, 2016) proposes a television program recommendation system designed to increase audience engagement and revitalize traditional broadcasting, particularly in light of competition from digital media. The system leverages Big Data analytics to deliver personalized recommendations based on viewers' watching habits. It employs data mining techniques to extract behavioral patterns and uses the Hadoop framework for large-scale data processing. Cross-analysis is performed across multiple parameters—such as program type, broadcast time, and audience size—to optimize recommendations.

The system architecture is divided into three main modules. Data are first collected via *Set-Top Boxes (STB)* and through online sources such as Baidu Billboard, Sohu News, Sina Weibo, and WeChat. These data are then analyzed and mined, and finally, the system generates personalized recommendations and scheduling suggestions based on popularity, peak viewing hours, and user profiles. This approach enables the broadcasting station to compete effectively with major Chinese networks (Dragon TV, Hunan Satellite TV, Jiangxi Satellite TV, and Zhejiang Satellite TV), improving both ratings performance and audience experience.

The article **"Research on E-Commerce Intelligent Recommendation System Based on Big Data Analysis"** (Yequn Huang, 2024) investigates how Big Data analytics can enhance intelligent recommendation systems in the e-commerce sector, focusing on personalization, market forecasting, and marketing optimization. The proposed model builds behavioral profiles using browsing history, keywords, purchases, and clicks. These data are processed through an error function *F* that compares expected and actual values in the recommendation model, thereby improving prediction accuracy. Additionally, the study employs a *network pruning* technique to eliminate less relevant neural connections, enhancing system efficiency. The resulting system improved marketing decision support, reduced the time between product viewing and purchase, and achieved superior performance in precision, click-through rate, and user satisfaction compared to other models.

The article **"Research on Multi-Dimensional Dynamic Recommendation Technology of Mobile E-Commerce Platform Based on Collaborative Filtering Algorithm"** (Shufang Zhao, 2023) presents a recommendation approach for mobile e-commerce platforms, aiming to enhance personalization and the efficiency of product recommendations. The system applies collaborative filtering techniques to improve recommendation accuracy, using dynamic algorithms such as *K-Nearest Neighbor (KNN)* to calculate user similarity. The evaluation focuses on measuring the proportion of products actually chosen by users that had been recommended by the system.

This model also incorporates the *Alternating Least Squares (ALS)* algorithm within the Spark MLlib library. To mitigate data sparsity and imprecise recommendations, the authors introduce a "double threshold" technique to distinguish between strong, weak, and irrelevant neighbors, thereby reducing the *Mean Absolute Error (MAE)* and improving overall system accuracy. The proposed model outperformed other algorithms in mobile e-commerce environments, where fast response times are essential to increasing user conversion rates.

The article **"Research on Recommendation Algorithm Based on E-commerce User Behavior Sequence"** (Bai Li, 2021) proposes a recommendation system for e-commerce platforms based on the chronological sequence of user interactions. The system integrates offline and real-time data mining techniques with deep learning models. Offline data mining is executed through the Hadoop platform using MapReduce, and results are stored in a database. For real-time analysis, the system employs message subscription mechanisms in Kafka clusters and gathers real-time statistics through Spark clusters using the *DeepLearning4j* framework. This framework optimizes the online learning process and generates recommendation outputs that are continuously updated in a real-time database. The model also uses tools such as Flume, Kafka, and Spark Streaming to collect, process, and analyze browsing, click, and purchase data in real time. The resulting system is robust, accurate, and adaptable, demonstrating strong performance in pattern extraction and high effectiveness in e-commerce environments.

The article **"Research on User Big Data Intelligent Recommendation Technology Supported by Mining Algorithm"** (Hulin Jin, 2024) explores a Big Data–driven recommendation approach based on data mining algorithms to address information overload on digital platforms. The system analyzes user behavior patterns—including browsing and search histories as well as social media activity—using *K-Means* clustering and *Apriori/FP-Growth* association rule algorithms. The K-Means algorithm groups users with similar characteristics, while the association rule algorithms identify frequent patterns of consumption and interest, which are then used to predict and suggest relevant content. The approach also integrates collaborative filtering, resulting in a hybrid recommendation model.

A case study conducted in this work demonstrated that the implemented system significantly improved information filtering efficiency and recommendation quality. The hybrid implementation reduced the time required to locate relevant content and enhanced the overall user experience in data-intensive environments.

DISCUSSION

After analyzing the five selected studies, the discussion is presented below by addressing the three previously defined research questions.

Data Collection (Q1).

In most of the reviewed papers, data collection is performed through users' browsing histories, clicks, and purchase records. This method appears in four of the five analyzed studies, all of which focus on e-commerce platforms. The study on television program recommendation by (Mengyi Zhang, 2016) differs by using web crawlers to automatically gather program schedule data, as well as a device located in users' homes to collect viewing information.

Tools and Technologies (Q2).

The analysis of the five studies reveals a wide diversity of tools and technologies used in the development of intelligent recommendation systems. In all cases, the selected tools reflect the need to handle large volumes of data, process information in real time, and provide highly personalized recommendations. Both (Shufang Zhao, 2023) and (Bai Li, 2021) adopt collaborative filtering methods; however, the former applies it as the core recommendation algorithm, while the latter combines it with content-based recommendations, resulting in a hybrid system.

In (Mengyi Zhang, 2016), the Hadoop framework is used for large-scale data processing and multiple regression analysis, supported by data mining algorithms to identify viewing patterns and preferences. (Yequn Huang, 2024) applies an error function based on fuzzy logic to compare expected and actual values, enhancing data accuracy, and employs *network pruning* to optimize the neural network by removing redundant connections. (Bai Li, 2021) focuses on real-time analytics, leveraging tools such as Flume, Kafka, and Spark Streaming for processing collected data. To enable model learning, deep learning algorithms implemented through the *Deeplearning4j* framework are used to identify behavioral patterns and predict relevant products for each user with high accuracy.

Across the analyzed studies, the tools employed range from Big Data processing platforms and deep learning frameworks to real-time analytics engines and traditional data mining algorithms. Despite their methodological diversity, all share the common goal of improving the accuracy, efficiency, and quality of recommendation systems.

Scalability and Performance (Q3).

Scalability and performance emerge as two of the most significant challenges faced by recommendation systems, as they must process large volumes of data in real time. In (Mengyi Zhang, 2016), these challenges are addressed through parallel data processing, ensuring consistent system performance. The study also employs efficient mining algorithms and multidimensional analysis—cross-referencing multiple data variables to identify complex behavioral patterns—allowing fast information processing and scalability as the user base grows.

In (Yequn Huang, 2024) the fuzzy logic–based error function filters and normalizes data before feeding it into the recommendation model, significantly reducing computational complexity and improving system performance. Scalability is achieved through *network pruning*, which removes redundant neural connections, enabling the system to operate more efficiently even with a high number of users and products.

(Shufang Zhao, 2023), addresses scalability through the Spark platform, enabling large-scale, high-speed data processing. The *Alternating Least Squares (ALS)* algorithm integrated into Spark MLlib allows the system to adapt quickly to changing user behavior while maintaining strong performance, even in high-traffic mobile environments. (Bai Li, 2021) deals with real-time data streams using a combination of Flume, Kafka, and Spark Streaming, which allows user behavior data to be collected and processed continuously without compromising performance. The use of the *Deeplearning4j* framework further enables parallel and scalable training of deep learning models.

Finally, (Hulin Jin, 2024) focuses on optimized data mining algorithms that remain efficient even when applied to massive datasets. The proposed system groups users into clusters and applies association rules to predict interests, achieving horizontal scalability without degrading performance. Additionally, by combining multiple recommendation techniques into a hybrid approach, the system balances computational load and improves accuracy without sacrificing speed.

The analyzed recommendation systems address scalability and performance challenges through distributed processing, algorithmic optimization, Big Data frameworks, and intelligent data filtering and compression methods. These strategies result in responsive, precise, and adaptive systems capable of maintaining high efficiency despite the continuous growth of user numbers and data volumes.

CONCLUSION

The analysis of Big Data–based recommendation systems, conducted through the review of five selected scientific articles published between 2015 and 2025, revealed that modern recommendation systems rely on Big Data technologies, data mining algorithms, and deep learning techniques to deliver more accurate and personalized results.

The findings show that data collection primarily occurs through user interactions with digital platforms, encompassing browsing histories, clicks, purchases, and search queries. Among the identified tools, frameworks such as Hadoop and Spark were the most prominent, alongside classical data mining algorithms like K-Means, Apriori, and FP-Growth, advanced deep learning approaches, and hybrid collaborative filtering methods. Each study proposed specific adaptations of these technologies to maximize recommendation performance and accuracy.

Regarding scalability and performance, the reviewed works presented solutions based on distributed processing, algorithmic optimization, large-scale parallel frameworks, and real-time analysis. These strategies enable recommendation systems to maintain high efficiency and responsiveness even as user bases and data volumes continue to grow.

This study demonstrates that Big Data–driven recommendation systems are essential for enhancing user-to-customer conversion and improving digital experiences. Their growing relevance underscores the importance of intelligent, data-driven personalization in an increasingly complex and data-intensive technological ecosystem.

REFRENCES

[1] Abbaschian, B. J. (2017). A Review of Hybrid Recommender Systems. AD ALTA: Journal of Interdisciplinary Research (p. 72). DOI 10.3353/ajir.07.02.19 .

[2] Bai Li, A. M. (2021). Research on Recommendation Algorithm Based on E-commerce User Behavior Sequence. 2021 IEEE 2nd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA). Guangzhou, China: 10.1109/ICIBA52610.2021.9688086, 978-1-6654-2877-4.

[3] Hulin Jin, X. S.-G. (2024). Research on User Big Data Intelligent Recommendation Technology Supported by Mining Algorithm. 2024 International Conference on Intelligent Computing and Next Generation Networks (ICNGN). Hefei, China: 10.1109/ICNGN63705.2024.10871642, 979-8-3315-2922-2.

[4] Kyung-Yong Jung, D.-H. P.-H. (2004). Hybrid Collaborative Filtering and Content-Based Filtering for Improved Recommender System. ICCS 2004 – LNCS 3036 (pp. 295-302). Bubak, M. et al. (Eds.): Springer-Verlag, Berlin Heidelberg.

[5] Mengyi Zhang, M. S. (2016). A TV Program Recommendation System Based on Big Data. 2016 International Conference on Industrial Informatics - Computing Technology, Intelligent Technology, Industrial Information Integration (ICIICII). Wuhan, China: 10.1109/ICIICII.2016.00128, 978-1-5090-0806-3.

[6] Page MJ, M. J. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. 372:n71. doi:10.1136/bmj.n71.: BMJ.

[7] Personalized Recommendations: How Netflix and Amazon Use Deep Learning to Enhance User Experience. (2025, 04 09). Retrieved from Medium: https://medium.com/@zhonghong9998/personalized-recommendations-how-netflix-and-amazon-use-deep-learning-to-enhance-user-experience-e7bd6fcd18ff

[8] Shufang Zhao, G. S. (2023). esearch on Multi-Dimensional Dynamic Recommendation Technology of Mobile E-Commerce Platform Based on Collaborative Filtering Algorithm. 2023 International Conference on Networking, Informatics and Computing (ICNETIC). Nanchang, China: 10.1109/ICNETIC59568.2023.00116, 979-8-3503-1331-4.

[9] Yequn Huang, S. Y. (2024). Research on E-Commerce Intelligent Recommendation System Based on Big Data Analysis. 4th International Conference on Electronic Information Engineering and Computer Technology (EIECT). Guangzhou, China: 10.1109/EIECT64462.2024.10866645, 979-8-3315-2885-0.