JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Survey on the Evolution of Recommender Systems: From Classical Models to LLM-based Approaches

[1]Bhakti Ahirwadkar, [2]Seema Chaudhary
[1]Associate Professor, [2] Associate Professor
[1]Computer Science & Engineering,
[1]Maharashtra Institute of Technology, Aurangabad, India

Abstract: Recommender systems (RS) have evolved from simple collaborative filtering models to sophisticated frameworks powered by large language models (LLMs). This survey reviews the trajectory of research over the past three decades, focusing on the shift from classical recommendation paradigms to deep learning and recent LLM-based methods. We classify the literature into six stages: early collaborative and content-based filtering, hybrid and ensemble models, context-aware systems, deep learning-based recommender systems, explainable recommendation, and the emergence of LLM-powered systems. Key milestones, representative approaches, and open challenges are highlighted, providing a comprehensive overview for researchers and practitioners.

***IndexTerms* – Recommender System, Large Language Model (LLM), Prompt Engineering, LLM Augmented RS, Deep Neural Network, CNN, Autoencoder, RNN**

I. INTRODUCTION

Recommender systems (RS) have become indispensable in modern digital ecosystems, from e-commerce and streaming platforms to healthcare and education. In an age of information abundance, RS serve as intelligent intermediaries that alleviate the problem of information overload by analyzing user behavior, inferring preferences, and predicting items of potential interest. From suggesting movies on Netflix and products on Amazon to recommending learning resources and personalized treatments, recommender systems now shape much of the user experience in digital environments. The field of RS has undergone many developments since the mid-1990s, when collaborative filtering was first introduced (P. Resnick et. al., 1994). Over the years, successive paradigms have built on and expanded earlier foundations: item-based approaches (B. Sarwar, 2001) which emphasis on current user profile and item features and hybrid techniques (R. Burke, 2002). With the increasing volume and heterogeneity of user–item data, ensemble and machine learning–driven methods were explored (Y. Koren et. al., 2006), combining multiple algorithms to enhance prediction accuracy and robustness. A major paradigm shift occurred in the mid-2010s with the advent of deep learning, which enabled recommender systems to automatically learn hierarchical feature representations from raw data (B. Hidasi et. al., 2015), (S. Sedhain et. al., 2015), (X. He et. al., 2018), (S. Sharma et. al., 2021). Deep neural networks, convolutional models, and autoencoders revolutionized RS architectures by capturing complex, nonlinear patterns in user–item interactions. These systems are pushing the field toward a new era of language-informed, interactive, and generalizable recommendation.

The latest wave of research explores how large language models (LLMs) can transform recommendation by leveraging natural language understanding, reasoning, and conversational interaction (P. Liu et. al., 2023), (Q. Liu et. al., 2025). LLM-based recommenders leverage textual semantics, knowledge grounding, and conversational interfaces to provide more human-like and explainable recommendations. This survey aims to provide a comprehensive review of RS development, synthesizing insights and covering this evolution.

2. Early Foundations of Recommender Systems

2.1 Collaborative Filtering

Collaborative filtering (CF) pioneered personalized recommendation by leveraging the wisdom of crowds. The basic assumption is that users with similar preferences in the past will continue to share preferences in the future. Memory-based CF methods included user-based (P. Resnick et. al., 1994) and item-based algorithms (B. Sarwar, 2001). These methods introduced key concepts such as similarity measures, neighborhood models, and rating prediction.

Despite their success, classical CF approaches struggle with data sparsity, cold-start problems, and scalability challenges as datasets are growing in size.

2.2 Content-Based Filtering

In parallel, content-based filtering (CBF) utilizes item features such as keywords, tags, or metadata to construct user profiles (M. Pazzani, 2005). Recommendations are generated by matching a user profile with candidate items. While effective in domains where rich item attributes are available, CBF often suffers from overspecialization and limited ability to capture user novelty-seeking behavior.

3. Hybrid and Ensemble Recommender Systems

3.1 Hybrid Models

To address the limitations of CF and CBF, hybrid approaches emerged in the early 2000s (R. Burke, 2002). These models combine multiple recommendation techniques — such as collaborative, content-based, and demographic features—to improve coverage and accuracy. Techniques include weighted hybridization, feature augmentation, and switching strategies.

3.2 Ensemble-Based Methods

By the late 2000s, ensemble methods gained prominence, drawing from advances in Machine Learning. The effectiveness of RS enhances by combining diverse models (e.g., matrix factorization, neighborhood models, restricted Boltzmann machines) into a single system (Y. Koren et. al., 2006). These approaches substantially boost accuracy but often at the expense of interpretability and computational cost.

4. Context-Aware and Sequential Recommender Systems

4.1 Context-Aware Recommendation

With the proliferation of mobile and ubiquitous computing, researchers incorporated contextual information such as time, location, device type, and social context into recommendation (G. Adomavicius et. al., 2011), (G. Adomavicius et. al., 2005). Context-aware recommender systems (CARS) expanded the scope beyond static user–item interactions, capturing situational dynamics and improving personalization in real-world applications.

4.2 Sequential and Session-Based RS

Another milestone was the development of sequential recommender systems, which model the temporal order of user interactions. Early techniques used Markov chains, later replaced by Recurrent Neural Networks (RNNs) (B. Hidasi et. al., 2015) and attention mechanisms. These methods excel at predicting the “next item” in sessions, crucial for domains like e-commerce and music streaming.

5. Deep Learning in Recommender Systems

5.1 Representation Learning

The Deep Learning revolution marked a transformative era in recommender systems, enabling models to learn complex, nonlinear relationships between users and items through hierarchical representation learning. Unlike traditional matrix factorization methods that rely on linear latent factors, deep architectures can automatically extract high-level abstractions from diverse data sources such as text, images, and user behavior logs. This paradigm shift has allowed RS to move from shallow feature engineering toward end-to-end learning frameworks that integrate feature extraction, interaction modeling, and prediction within a unified architecture.

Among the pioneering deep learning approaches, Autoencoders for Collaborative Filtering represented one of the earliest successful integrations of neural networks with CF. By reconstructing partially observed user–item matrices, Autoencoders learn latent representations that effectively capture hidden user preferences and item characteristics, achieving superior generalization and robustness to data sparsity (S. Sedhain et. al., 2015), (F. Strub, 2018), (Y. Wu, 2016).

Convolutional Neural Networks (CNNs) (S. Sharma at. al., 2021) have also played a pivotal role in enhancing recommendation performance, particularly in domains where visual or textual features are influential. CNN-based models can automatically extract discriminative patterns from product images, movie posters, or item descriptions, thereby bridging the gap between content-based and collaborative signals (Y. Gong, 2016), (X. Wang, 2017). This capability has been crucial in applications like fashion, multimedia, and e-commerce recommendations.

Similarly, Recurrent Neural Networks (RNNs) have been widely adopted to model sequential and temporal dynamics in user interactions (B. Hidasi et. al., 2015), (Y. Ko et. al., 2016). By capturing order-sensitive behavioral patterns such as browsing history or session-based clicks RNN-based recommenders can predict a user’s next likely action or preference with high contextual relevance. An attention-based LSTM for recommendation systems enhances personalization by allowing the Long Short-Term Memory (LSTM) network to dynamically weigh the importance of different user interactions over time, thus capturing complex and long-term user preferences to generate more relevant suggestions (Y. Li et. al., 2016). This has proven especially effective in personalized news feeds and session-based recommendations.

Finally, Multi-Layer Perceptrons (MLPs) (X. He et. al., 2018) introduced a flexible framework for nonlinear interaction modeling between users and items. By learning complex mappings from user and item embeddings to preference scores, MLP-based approaches (such as Neural Collaborative Filtering) overcome the linearity limitations of traditional MF models and enable more expressive modeling of user–item relationships.

Collectively, these deep learning architectures have significantly advanced the predictive accuracy and personalization capability of recommender systems. However, they also come with notable challenges, including the need for large-scale annotated datasets, high computational resources, and careful hyperparameter tuning to avoid overfitting. Despite these limitations, deep neural models continue to form the core backbone of modern recommendation research, providing a strong foundation upon which newer paradigms such as graph-based learning and LLM-driven recommendation are being built.

5.2 Transformer-Based Models

More recently, transformers have emerged as a unified architecture for RS, leveraging self-attention to model complex dependencies (F. Sun et. al., 2019). They outperform traditional deep models in sequential and multimodal recommendation tasks.

6. Explainable and Trustworthy Recommender Systems

As RS permeated critical applications, explainability became essential for user trust and system transparency. Explainable recommender systems focus on generating human-understandable rationales behind recommendations (Y. Zhang et. al., 2018). Methods range from attention-based visualization to natural language explanations. This aligns RS research with ethical AI principles, emphasizing fairness, accountability, and interpretability.

7. Large Language Models for Recommendation

7.1 Prompt-Based Recommendation

The advent of GPT-style Large Language Models (LLMs) has fundamentally reshaped the recommender system (RS) landscape, introducing a new paradigm that integrates language understanding, reasoning, and generation into the recommendation process. Unlike traditional RS models that depend on structured user–item matrices, LLM-based recommenders can process unstructured natural language inputs such as reviews, descriptions, and conversational queries to infer user intent (P. Liu et. al., 2023), (L. Xu et. al., 2018), (S. Joshna et. al., 2023).

Through the mechanism of prompt engineering, users can interact with recommendation models in free-form dialogue, expressing preferences, constraints, or contextual cues in natural language. The LLM interprets these signals to generate personalized recommendations directly, often without explicit training on a specific dataset. This zero-shot and few-shot capability allows LLMs to generalize across domains and adapt to novel contexts with minimal supervision.

Furthermore, prompt-based approaches facilitate interactive and conversational recommendation, where the system engages in multi-turn dialogue to clarify preferences, justify suggestions, and adapt its output dynamically. This marks a significant departure from static recommendation pipelines toward human-centered, explainable, and context-aware systems. In essence, prompt-based recommendation bridges the gap between human language and machine reasoning, enabling recommendation to become more transparent, adaptive, and user-aligned.

7.2 LLM-Augmented Architectures

Building upon prompt-based capabilities, recent studies have proposed LLM-augmented recommender architectures that integrate language models into the core RS pipeline (Q. Liu et. al., 2025). In such hybrid frameworks, LLMs function as controllers or meta-reasoners, coordinating multiple components such as candidate retrieval, ranking, filtering, and explanation generation. Instead of replacing traditional RS modules, the LLM operates as a high-level decision engine that interprets user context, issues sub-queries to retrieval systems, and synthesizes final recommendations into coherent, natural language outputs.

These architectures often employ retrieval-augmented generation (RAG) (S. Wang et. al., 2025) or modular hybrid designs, where the LLM collaborates with embedding-based retrieval models or graph-based recommenders. The LLM provides interpretability and contextual reasoning, while the underlying RS components ensure efficiency and accuracy (W. Wei et. al., 2023,), (D. Nawara et. al., 2025).

An additional advancement involves feedback loops that enable continuous personalization. As users interact conversationally accepting, rejecting, or refining recommendations the LLM incorporates these signals to adapt its prompts or fine-tune its responses in real time. This dynamic learning process enhances personalization and fosters a more interactive, lifelong learning paradigm in recommender systems.

Overall, LLM-augmented architectures represent a promising synthesis of neural recommendation and natural language intelligence, pushing the field toward generalist AI recommenders capable of contextual reasoning, transparency, and continuous adaptation.

7.3 Opportunities and Challenges

While promising, LLM-based RS faces challenges including:

- High computational and deployment costs.
- Limited interpretability of LLM reasoning.
- Data privacy and hallucination risks.
- Integration with structured recommendation pipelines.

8. Open Challenges and Future Directions

Despite rapid progress, several open problems remain:

- Scalability: Efficiently handling billion-scale datasets.
- Fairness and Bias Mitigation: Ensuring equitable recommendations across demographics.
- Multimodal Fusion: Integrating text, images, audio, and behavioral data.
- Explainability vs. Performance: Balancing accuracy with transparency.
- Sustainable RS: Reducing the environmental cost of large models.
- Conversational and LLM-driven RS: Advancing human-like, interactive systems while addressing ethical concerns.

9. Conclusion

Recommender systems have undergone a profound development over three decades, evolving from neighborhood-based collaborative filtering to hybrid systems, deep learning, and now LLM-powered frameworks. Each paradigm has contributed unique insights and capabilities, yet unresolved challenges continue to inspire future research. With the convergence of deep learning, natural language processing, and generative AI, RS stands at the frontier of human-centered AI, poised to play an even greater role in shaping personalized digital experiences.

References

[1] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom and John Riedl, "GroupLens: An open architecture for collaborative filtering of netnews", 1994, ACM.

[2] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl "Item-based collaborative filtering recommendation algorithms", WWW, 2001, ACM.

[3] R. Burke, "Hybrid recommender systems: Survey and experiments". UMUAI, 2002.

[4] Gediminas Adomavicius and Alexander Tuzhilin "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions" IEEE TKDE, June 2005.

[5] Michael J. Pazzani and Daniel Billsus, "Content-Based Recommendation Systems", The Adaptive Web, LNCS 4321, pp. 325 – 341, 2007.

[6] Yehuda Koren, Robert Bell and Chris Volinsky , "Matrix factorization techniques for recommender systems", IEEE Computer 2009.

[7] Gediminas Adomavicius, Bamshad Mobasher, Francesco Ricci, and Alex Tuzhilin, "Context-Aware Recommender Systems", AI Magazine, Vol 32, No. 3, 2011.

[8] Balazs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, Domonkos Tikk, "Session-based recommendations with recurrent neural networks" ICLR 2015.

[9] Suvash Sedhain, Aditya Krishna Menon, Scott Sanner and Lexing Xie, "AutoRec: Autoencoders Meet Collaborative Filtering", WWW 2015.

[10] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, Tat-Seng Chua, "Neural collaborative filtering", WWW 2017.

[11] Yongfeng Zhang and Xu Chen, "Explainable recommendation: A survey and new Perspectives" IJCAI 2018.

[12] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou and Peng Jiang "BERT4Rec: Sequential recommendation with bidirectional encoder representations from Transformer". CIKM 2019.

[13] Sunny Sharma, Vijay Rana and Vivek Kumar, "Deep learning based semantic personalized recommendation system", International Journal of Information Management Data Insights, Volume 1, Issue 2, November 2021.

[14] Peng Liu, Lemei Zhang and Jon Atle Gulla, "Pre-train, Prompt, and Recommendation: A Comprehensive Survey of Language Modeling Paradigm Adaptations in Recommender Systems", Transactions of the Association for Computational Linguistics, vol. 11, pp. 1553–1571, 2023.

[15] Qidong Liu, Xiangyu Zhao, Yuhao Wang, Yejing Wang, Zijian Zhang, Yuqi Sun, Xiang Li, Maolin Wang, Pengyue Jia, Chong Chen, Wei Huang and Feng Tian, "Large Language Model Enhanced Recommender Systems: Methods, Applications and Trends", KDD 2025.

[16] Yuyun Gong and Qi Zhang, "Hashtag Recommendation Using Attention-Based Convolutional Neural Network", Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16).

[17] Xuejian Wang, Lantao Yu, Kan Ren, Guanyu Tao, Weinan Zhang, Yong Yu and Jun Wang "Dynamic A,ention Deep Model for Article Recommendation by Learning Human Editors' Demonstration", KDD 2017.

[18] Florian Strub, Romaric Gaudel and Jeremie Mary, "Hybrid Recommender System based on Autoencoders" the 1st Workshop on Deep Learning for Recommender Systems, Sep 2016, Boston, United States. pp.11- 16, 2016,

[19] Yao Wu, Christopher DuBois, Alice X. Zheng and Martin Ester, "Collaborative Denoising Auto-Encoders for Top-N Recommender Systems", WSDM'16, February 22–25, 2016.

[20] Yang Li, Ting Liu, Jing Jiang and Liang Zhang, "Hashtag Recommendation with Topical Attention-Based LSTM", Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics.

[21] Young-Jun Ko, Lucas Maystre and Matthias Grossglauser, "Collaborative Recurrent Neural Networks for Dynamic Recommender Systems", JMLR: Workshop and Conference Proceedings 63:366381, 2016.

[22] Lanling Xu, Junjie Zhang, Bingqian Li, Jinpeng Wang, Mingchen Cai, Wayne Xin Zhao, and Ji-Rong Wen, "Prompting Large Language Models for Recommender Systems: A Comprehensive Framework and Empirical Analysis" In . ACM 2018

[23] Sarabu Joshna and B.Ramya Sree, "Scientific Research Paper Recommendation Using GPT-3 Prompt Engineering", JETIR November 2023, Volume 10, Issue 11.

[24] Shijie Wang, Wenqi Fan, Yue Feng, Shanru Lin, Xinyu Ma, Shuaiqiang Wang and Dawei Yin "Knowledge Graph Retrieval-Augmented Generation for LLM-based Recommendation", Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics 2025.

[25]Wei Wei, Xubin Ren, Jiabin Tang, Qinyong Wang, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin and Chao Huang, "LLMRec: Large Language Models with Graph Augmentation for Recommendation", Proceedings of the 17th ACM International Conference on Web Search and Data Mining

[26] Dina Nawara and Rasha Kashef, "A Comprehensive Survey on LLM-Powered Recommender Systems: From Discriminative, Generative to Multi-Modal Paradigms", IEEE Access, Vol. 13, 2025.