A Survey on Modern Recommender Systems: Collaborative Filtering to LLM-based Recommender System

Bonwoo Koo
Dept. of Industrial & Systems Engineering
KAIST
Daejeon, South Korea
kbw9896@kaist.ac.kr

***Abstract*— This survey explores the evolution of recommender systems from early collaborative filtering methods to advanced multi-modal Large Language Model (LLM)-based approaches. Key categories discussed include collaborative filtering, hybrid models, graph-based methods, session-based approaches, LLM-based systems, and multi-modal LLM-based models. Despite advancements, challenges such as cold start problems, data sparsity, scalability, computational complexity, and modality gaps remain. Future work should improve the efficiency and scalability of LLM-based models, extend their applicability to various data types, and enhance performance across diverse domains with a unified approach. These improvements are essential for implementing a unified, end-to-end, multi-task recommendation model suitable for real-world usage and industry applications.**

Keywords*— *Recommender System, Collaborative Filtering, Large Language Models, Multimodal Recommendation

I. Introduction

Recommender systems have become an integral part of modern digital ecosystems, significantly influencing user experiences across a wide array of domains, including e-commerce, social media, entertainment, and online content platforms. Their primary function is to filter and present personalized information from vast datasets, thereby helping users navigate the overwhelming abundance of choices. Over the past few decades, the field of recommender systems has witnessed remarkable advancements, evolving through various paradigms and methodologies.

The field of recommender systems began with **content-based filtering** techniques, which focus on the attributes of items and users. Content-based filtering recommends items by analyzing the features of items that a user has previously interacted with and suggesting similar items. This approach, while useful in handling new items, often suffers from limited novelty in recommendations and an inability to capture complex user preferences that go beyond observed behavior.

To address some of the limitations of content-based filtering, **collaborative filtering** emerged. Over time, it became the predominant algorithm in modern recommender systems, forming the foundation for many subsequent advancements in the field. Collaborative filtering leverages user-item interaction data to make recommendations, predicting what a user might like based on the preferences of similar users or items. By identifying patterns in user behavior, collaborative filtering can provide more diverse and accurate recommendations. However, it also faces challenges such as data sparsity and the cold-start problem, where insufficient data on new users or items hampers the systemʹs ability to provide accurate recommendations.

The subsequent evolution led to the development of **hybrid recommender systems**, which combine collaborative and content-based methods to harness the strengths of both approaches while mitigating their respective weaknesses. These systems employ various strategies, such as model blending and ensemble techniques, to enhance recommendation accuracy and robustness.

As the complexity and scale of data increased, **graph-based recommender systems** were introduced to model intricate relationships between users and items. Utilizing graph theory and network analysis, these systems capture higher-order connections and dependencies, enabling more nuanced and accurate recommendations.

The advent of **session-based recommender systems** marked a significant shift towards capturing temporal dynamics and sequential patterns in user behavior. By focusing on user interactions within a session, these models can provide highly relevant recommendations in real-time, catering to immediate user needs and preferences.

More recently, the rise of large language models (LLMs) has opened new frontiers in recommendation research. **LLM-based recommender systems** leverage the powerful contextual understanding and generative capabilities of these models to enhance recommendation quality, particularly in handling textual data and complex user intents.

Building upon this, **multi-modal LLM-based recommender systems** integrate various data modalities, such as text, images, and audio, to provide a richer and more holistic understanding of user preferences. By combining multiple sources of information, these systems can deliver highly personalized and contextually aware recommendations.

This survey paper aims to provide a comprehensive overview of the historical development of recommender systems, tracing the evolution of methodologies from collaborative filtering to the latest advances in multi-modal LLM-based approaches. By examining the strengths and limitations of each category, we seek to illuminate the progression of the field and identify emerging trends and future directions in recommender system research.

II. Collaborative Filtering based Recommender Systems

MF: Koren et al. [8] introduced Matrix Factorization (MF) as a foundational technique in collaborative filtering. MF decomposes the user-item interaction matrix into two lower-dimensional matrices, representing latent factors for users and items. By approximating the original interaction matrix through the product of these two matrices, MF captures underlying patterns and relationships in the data. This approach enables the prediction of missing entries in the interaction matrix, thereby providing recommendations. MF is known for its scalability and effectiveness in capturing user preferences and item characteristics, making it a widely adopted method in recommender systems.

PMF: Mnih and Salakhutdinov [11] introduced Probabilistic Matrix Factorization (PMF) to enhance traditional matrix factorization by incorporating a probabilistic framework. PMF models the user-item interaction matrix as a probabilistic distribution, assuming Gaussian noise on the observed interactions. This probabilistic approach allows for a principled handling of uncertainty and missing data. By optimizing the likelihood of the observed data, PMF learns latent factors for users and items that best explain the interactions. This method provides a robust and flexible framework for recommendation, especially in scenarios with sparse data.

OCCF: Pan et al. [12] proposed One-Class Collaborative Filtering which addresses the challenge of recommending with only positive feedback, such as clicks or purchases, without explicit negative feedback. It is a method that adapts matrix factorization techniques to handle implicit feedback. By treating unobserved interactions as missing data or by generating pseudo-negative examples, OCCF models can infer user preferences from the available positive interactions. This approach is particularly useful in real-world applications where obtaining explicit ratings is difficult, enabling effective recommendations based on implicit user behavior.

BPR: Rendle et al. [14] introduced Bayesian Personalized Ranking (BPR) as a pairwise ranking optimization technique for collaborative filtering. BPR focuses on optimizing the ranking of items for each user by leveraging implicit feedback. The key idea is to learn a preference ranking from user interactions, assuming that users prefer observed items over unobserved ones. BPR optimizes a ranking loss function, typically using stochastic gradient descent, to learn latent factors that capture user preferences. This approach effectively addresses the challenges of implicit feedback and provides high-quality personalized recommendations.

FM: Rendle [13] introduced Factorization Machines (FM) as a generalization of matrix factorization techniques, capable of modeling higher-order interactions between variables. FMs combine the strengths of matrix factorization with the flexibility to incorporate additional features, such as user and item attributes. By factorizing the interactions between all pairs of features, FMs can capture complex patterns in the data. This makes FMs suitable for a wide range of recommendation tasks, including those involving side information. Their ability to handle high-dimensional sparse data and integrate various types of features has made FMs a versatile tool in recommender systems

III. Hybrid Recommender Systems

WD: Cheng et al. [2] introduced the Wide & Deep learning model, designed to combine the benefits of memorization and generalization in recommender systems. The model consists of two components: the wide part and the deep part. The wide part is a linear model that captures feature interactions explicitly, suitable for memorizing frequent co-occurrences. The deep part is a feedforward neural network that learns high-dimensional, non-linear interactions, enabling generalization to unseen feature combinations. By jointly training both components, the Wide & Deep model effectively leverages both shallow and deep features, enhancing recommendation accuracy and coverage.

DeepFM: Guo et al. [4] proposed DeepFM, an innovative model that seamlessly integrates the strengths of Factorization Machines (FM) and deep neural networks. The model consists of two components: an FM component for capturing low-order feature interactions and a deep neural network for modeling high-order interactions. Unlike other hybrid models, DeepFM shares the same input for both components, ensuring that both low-order and high-order feature interactions are learned from the same set of raw features. This approach not only captures intricate patterns in the data but also improves the model's ability to generalize, making it highly effective for recommendation tasks.

AutoRec: Sedhain et al. introduced AutoRec, an autoencoder-based collaborative filtering model. AutoRec employs autoencoders to reconstruct the user-item interaction matrix, learning latent representations of users and items in the process. The model consists of an encoder that compresses the interaction data into a lower-dimensional space and a decoder that reconstructs the original data from these latent representations. By minimizing the reconstruction error, AutoRec captures the underlying patterns in the interaction data. This method provides a robust framework for recommendation, leveraging the strengths of both collaborative filtering and deep learning.

IV. Graph-based recommender systems

GCN: Kipf and Welling [7] introduced Graph Convolutional Networks (GCN) as a powerful method for learning representations on graph-structured data. GCNs extend the concept of convolutional neural networks to graphs, enabling the aggregation of feature information from a node's local neighborhood. In the context of recommender systems, GCNs can effectively model the complex relationships between users and items by considering the connections in the user-item interaction graph. By stacking multiple graph convolutional layers, GCNs capture both local and global structural information, resulting in improved recommendation accuracy through enhanced representation learning.

GraphSAGE: Hamilton et al. [5] proposed GraphSAGE (Graph Sample and Aggregate) as an inductive framework for generating node embeddings in large graphs. Unlike traditional GCNs, GraphSAGE samples and aggregates features from a node's local neighborhood to generate embeddings, which allows it to generalize to unseen nodes. This inductive capability is particularly useful in recommender systems for handling dynamic graphs where new users or items are continuously added. By learning an aggregation function that combines node features, GraphSAGE can efficiently scale to large graphs, making it suitable for real-time recommendation scenarios.

GAT: Veličković et al. [18] introduced Graph Attention Networks (GAT), which leverage attention mechanisms to dynamically weigh the importance of a node's neighbors during feature aggregation. GATs assign different attention coefficients to each neighbor, allowing the model to focus on the most relevant nodes in the graph. This attention mechanism enhances the ability to capture complex and heterogeneous relationships in user-item interaction graphs. In recommender systems, GATs can provide more accurate recommendations by adaptively learning the importance of various interactions, resulting in a more fine-grained and personalized representation of user preferences.

V. Session-based recommender systems

SASRec: Kang and McAuley [6] introduced SASRec (Self-Attentive Sequential Recommendation), a model that utilizes self-attention mechanisms to capture sequential patterns in user behavior. SASRec leverages the Transformer architecture to model the temporal dependencies in user-item interactions within a session. By applying self-attention, SASRec can dynamically focus on relevant past interactions when predicting the next item, effectively capturing long-range dependencies and user preferences. This approach allows for highly accurate and context-aware recommendations, addressing the limitations of traditional sequential models that struggle with long-term dependencies.

BERT4Rec: Sun et al. [16] proposed BERT4Rec, a model that adapts the Bidirectional Encoder Representations from Transformers (BERT) for sequential recommendation tasks. BERT4Rec employs a bidirectional self-attention mechanism to model the contextual relationships between items in a user's interaction sequence. Unlike unidirectional models, BERT4Rec predicts masked items within the sequence, capturing the full context of user behavior. This bidirectional approach enables a deeper understanding of item dependencies and user preferences, resulting in more accurate and comprehensive recommendations. BERT4Rec's ability to leverage bidirectional context makes it particularly effective in scenarios where understanding the order and context of interactions is crucial.

VI. LLM-based Recommender systems

CoLLM: Zhang et al. [20] introduced LLM-based recommendation model which emphasizes the integration of collaborative information modeling with text semantics for recommendation systems. CoLLM utilizes external traditional models to seamlessly incorporate collaborative details into large language models. This integration enhances recommendation performance in both cold-start and warm-start scenarios, leveraging the strengths of both collaborative filtering and language models for improved accuracy [17].

TALLRec: Bao et al. [1] introduced TALLRec, an efficient and effective fine-tuning framework designed to optimize Large Language Models (LLMs) in recommendation task. TALLRec combines instruction tuning and recommendation tuning to enhance overall model effectiveness. The framework employs two tuning stages: instruct-tuning, which focuses on generalization using self-instruct data from Stanford Alpaca, and rec-tuning, which structures limited user interactions into recommendation instructions. This dual-stage tuning process allows LLM to gain robust cross-domain generalization and significantly improves the model's performance in recommendation tasks [17].

LLaRA: Liao et al. [9] introduced LLaRA, a sequential recommendation framework for Large Language Models (LLMs). LLaRA integrates ID-based item embeddings from traditional recommender systems with textual item features in LLM prompts. By treating "sequential user behavior" as a unique modality, an adapter facilitates the transition between traditional ID embeddings and the LLM input space. Through the use of curriculum learning, the training process gradually shifts from text-only prompting to a hybrid of text and item embeddings, enabling LLMs to proficiently manage sequential recommendation tasks [17].

VII. Multi-modal LLM-based recommender systems

VIP5: Geng et al. [3] proposed a parameter-efficient multimodal foundation recommendation model, named

VIP5 (Visual P5), under the P5 recommendation paradigm. P5 was the first unified framework that integrates various recommendation task and VIP5 extends this framework to a multi-modal setting, aiming to integrate visual, textual, and personalization modalities within a single architecture. To accomplish this, VIP5 introduces multimodal personalized prompts that can accommodate multiple modalities under a unified format. Additionally, VIP5 proposes a parameter-efficient training method for foundation models, which involves freezing the P5 backbone and fine-tuning lightweight adapters in attention blocks. This approach not only enhances recommendation across diverse modalities but also increases efficiency in terms of training time and memory usage.

UniMP: Wei et al. [19] introduced a Unified paradigm for Multi-modal Personalization systems (UniMP) to harness multi-modal data effectively and simplify the complexities of task- and modality-specific customization. UniMP aims to handle a wide range of personalized needs, including item recommendation, product search, preference prediction, explanation generation, and user-guided image generation. UniMP constructs a standardized data format that seamlessly integrates diverse user historical data and employs a cross-attention mechanism to facilitate multi-modal user modeling. It unifies several personalization tasks within a coherent token generation framework and incorporates context reconstruction and token-level reweighting to ensure alignment. This framework overcomes several key issues from existing models like VIP5, which fails to fully harness the potential of raw data, does not effectively capture interactions among different data types, and lacks the flexibility to handle diverse input and output requirements inherent in multi-task learning. Experimental results demonstrate that UniMP outperforms competitive baselines across various benchmark tasks, showcasing its enhanced performance.

Rec-GPT4V: Liu et al. [10] proposed Rec-GPT4V: Visual-Summary Thought (VST), a novel reasoning framework for leveraging large vision-language models (LVLMs) in multimodal recommendation systems. Rec-GPT4V addresses two key challenges: the lack of user preference knowledge and the difficulty in handling multiple image dynamics. To tackle the first challenge, the framework uses user history as in-context indicators of user preferences. For the second challenge, it prompts LVLMs to generate summaries of item images and utilizes image comprehension in the natural language space, combined with item titles, to evaluate user preferences for candidate items. Comprehensive experiments were conducted across four datasets using three LVLMs: GPT4-V, LLaVa7b, and LLaVa-13b. The results demonstrate the effectiveness of the VST approach in enhancing multimodal recommendation performance compared to other reasoning strategies such as in-context learning and chain-of-thought.

VIII. FUTURE WORK

Despite improvements in addressing cold start problems and enhancing explainability in various recommendation tasks through the integration of deep learning and language models with conventional recommender systems, several limitations remain. Cold start and data sparsity issues persist, depending on the data available to the recommendation model. Additionally, scalability and complexity issues become more pronounced as data size increases. This is further exacerbated by longer inference times and heavy computational requirements when using transformer-based or LLM-based recommendation models, making them infeasible for real-world industrial applications. Although LLM-based models have introduced explainable recommendation through various natural language tasks, they still suffer from a modality gap between user/item data in the LLM embedding space. Furthermore, these models depend heavily on the diversity and quality of training data, along with the training strategies used, such as prompting methods, which can result in bias and fairness issues. Future work should aim to improve both efficiency and accuracy of LLM-based recommendation models, extend their modality to a broader range of data types such as video and audio, and enhance their performance across diverse domains with a single end-to-end model.

IX. CONCLUSION

The evolution of recommender systems demonstrates a continuous effort to overcome the limitations of earlier models and enhance recommendation accuracy and applicability. Content-based filtering initiated this journey by focusing on item and user attributes, but its limitations in novelty and complexity led to the emergence of collaborative filtering, which utilized user-item interaction data for more diverse recommendations. To address data sparsity and cold-start issues in collaborative filtering, hybrid models combined content-based and collaborative methods, improving robustness and accuracy. As data complexity grew, graph-based systems leveraged graph theory to capture intricate user-item relationships, while session-based models focused on real-time, context-aware recommendations by analyzing temporal and sequential user interactions. The introduction of large language models (LLMs) brought powerful contextual understanding and generative capabilities, particularly for textual data, and multi-modal LLM-based systems further enriched recommendations by integrating text, images, and audio. This comprehensive exploration highlights their current impact and sets the stage for future advancements in recommender systems.

REFERENCES

[1] Bao, K., Zhang, J., Zhang, Y., Wang, W., Feng, F., & He, X. (2023, September). Tallrec: An effective and efficient tuning framework to align large language model with recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems* (pp. 1007-1014).

[2] Cheng, H. T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., ... & Shah, H. (2016, September). Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems* (pp. 7-10).

[3] Geng, S., Tan, J., Liu, S., Fu, Z., & Zhang, Y. (2023). Vip5: Towards multimodal foundation models for recommendation. *arXiv preprint arXiv:2305.14302*.

[4] Guo, H., Tang, R., Ye, Y., Li, Z., & He, X. (2017). DeepFM: a factorization-machine based neural network for CTR prediction. *arXiv preprint arXiv:1703.04247*.

[5] Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in neural information processing systems*, *30*.

[6] Kang, W. C., & McAuley, J. (2018, November). Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)* (pp. 197-206). IEEE.

[7] Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

[8] Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, *42*(8), 30-37.

[9] Liao, J., Li, S., Yang, Z., Wu, J., Yuan, Y., Wang, X., & He, X. (2023). Llara: Aligning large language models with sequential recommenders. *arXiv preprint arXiv:2312.02445*.

[10] Liu, Y., Wang, Y., Sun, L., & Yu, P. S. (2024). Rec-GPT4V: Multimodal Recommendation with Large Vision-Language Models. *arXiv preprint arXiv:2402.08670*.

[11] Mnih, A., & Salakhutdinov, R. R. (2007). Probabilistic matrix factorization. *Advances in neural information processing systems*, *20*.

[12] Pan, R., Zhou, Y., Cao, B., Liu, N. N., Lukose, R., Scholz, M., & Yang, Q. (2008, December). One-class collaborative filtering. In *2008 Eighth IEEE international conference on data mining* (pp. 502-511). IEEE.

[13] Rendle, S. (2010, December). Factorization machines. In *2010 IEEE International conference on data mining* (pp. 995-1000). IEEE.

[14] Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2012). BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*.

[15] Sedhain, S., Menon, A. K., Sanner, S., & Xie, L. (2015, May). Autorec: Autoencoders meet collaborative filtering. In *Proceedings of the 24th international conference on World Wide Web* (pp. 111-112).

[16] Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., & Jiang, P. (2019, November). BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management* (pp. 1441-1450).

[17] Vats, A., Jain, V., Raja, R., & Chadha, A. (2024). Exploring the Impact of Large Language Models on Recommender Systems: An Extensive Review. *arXiv preprint arXiv:2402.18590*.

[18] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*.

[19] Wei, T., Jin, B., Li, R., Zeng, H., Wang, Z., Sun, J., ... & Tang, X. (2023, October). Towards Universal Multi-Modal Personalization: A Language Model Empowered Generative Paradigm. In *The Twelfth International Conference on Learning Representations*.

[20] Zhang, Y., Feng, F., Zhang, J., Bao, K., Wang, Q., & He, X. (2023). Collm: Integrating collaborative embeddings into large language models for recommendation. *arXiv preprint arXiv:2310.19488*.