Advancements in Modern Recommender Systems: Industrial Applications in Social Media, E-commerce, Entertainment, and Beyond

Sankalp KJ, Sai Naveena BV, Charith Chandra Sai Balne, Vinodh Kumar Sunkara, Sreyoshi Bhaduri, Vinija Jain, Aman Chadha

Advancements in Modern Recommender Systems: Industrial Applications in Social Media, E-commerce, Entertainment, and Beyond

Sankalp KJ[1], Sai Naveena BV[2] Charith Chandra Sai Balne[3]
Vinodh Kumar Sunkara[5], Sreyoshi Bhaduri[4]*, Vinija Jain[6], Aman Chadha[7]*
[1]Artificial Intelligence Institute, University of South Carolina
[2]Illinois Institute of Technology [3]University of Southern California
[4]Amazon [5]Meta [6]Stanford University [7]Amazon GenAI
sjajee@email.sc.edu, hi@vinija.ai, hi@aman.ai

Abstract

In the current digital era, the proliferation of online content has overwhelmed users with vast amounts of information, necessitating effective filtering mechanisms. Recommender systems have become indispensable in addressing this challenge, tailoring content to individual preferences and significantly enhancing user experience. This paper delves into the latest advancements in recommender systems, analyzing 115 research papers and 10 articles, and dissecting their application across various domains such as e-commerce, entertainment, and social media. We categorize these systems into content-based, collaborative, and hybrid approaches, scrutinizing their methodologies and performance. Despite their transformative impact, recommender systems grapple with persistent issues like scalability, cold-start problems, and data sparsity. Our comprehensive review not only maps the current landscape of recommender system research but also identifies critical gaps and future directions. By offering a detailed analysis of datasets, simulation platforms, and evaluation metrics, we provide a robust foundation for developing next-generation recommender systems poised to deliver more accurate, efficient, and personalized user experiences, inspiring innovative solutions to drive forward the evolution of recommender technology.

1 Introduction

The explosion of digital content and user-generated information in recent years has revolutionized how individuals interact with online platforms. However, this abundance of data has also introduced the significant challenge of information overload, making it difficult for users to find relevant content efficiently. Recommender systems have emerged as a critical solution to this problem, leveraging sophisticated algorithms to filter and personalize content, thereby enhancing user experience and reducing search time. Recommender systems are currently employed across various domains, including e-commerce, social media, entertainment, and healthcare, to suggest products, services, and information tailored to individual preferences. Despite their widespread application and success, these systems face persistent challenges such as scalability, cold-start issues, and data sparsity, which hinder their performance and effectiveness.

In industry, most recommendation systems use a two-stage approach (Covington et al., 2016): L1 (Candidate Generation) and L2 (Ranking). The first stage quickly narrows down a vast pool of items to a manageable subset of relevant candidates using efficient, scalable algorithms like collaborative filtering, two-tower models etc. The second stage precisely ranks the candidates with complex models, often deep neural networks, incorporating rich features to deliver a refined list of top recommendations tailored to user preferences. At the crux of it, using the right architecture that fits a specific recommendation need is crucial to achieving optimal performance and user satisfaction, essential for successful deployment of recommendation systems.

This paper provides a comprehensive review of 115 papers and 10 articles showcases recent advancements in recommender systems, categorizing them into content-based, collaborative, and hybrid approaches. It examines the evolution of these systems, evaluates the datasets and simulation platforms used in research, and discusses key performance metrics. By identifying existing gaps and challenges, this review aims to offer insights for future research and development of more efficient and versatile recommender systems, ultimately striving to improve user satisfaction across diverse applications.

*Work does not relate to position at Amazon.

2 Methodology

The research on recommender systems has seen substantial growth, with notable advancements across various domains such as e-commerce, entertainment, and social media. To provide a comprehensive overview of this dynamic field, we conducted an extensive review of technical papers on recommender systems.

2.1 Search Strategy

Our search strategy encompassed multiple databases including Google Scholar, IEEE Xplore, ACM Digital Library, and arXiv. We used a combination of keywords and phrases such as "Social media recommender systems", "Hybrid recommender systems in entertainment", "Health behavior predictions", "Financial recommender systems", and "personalization" to identify relevant studies. Additionally, we manually screened the references of the selected papers to identify any additional relevant studies that may have been missed during the initial search.

2.2 Inclusion and Exclusion Criteria

To ensure the quality and relevance of the studies included in our review, we established predefined inclusion and exclusion criteria. Studies were included if they:

- Focused on recommender systems.
- Provided empirical results or theoretical contributions to the field.

Studies were excluded if they:

- Were not available in full text.
- Were not written in English.
- Did not provide sufficient detail on the methodologies or results.

2.3 Data Extraction and Synthesis

After applying the inclusion and exclusion criteria, we screened the titles and abstracts of over 500 papers, yielding 57 studies that were included in our final review. For each included study, we extracted data on the following aspects:

- The specific domains or applications addressed.
- The type of recommender models used.

2.4 Analysis and Taxonomy Development

We conducted a qualitative and quantitative analysis of the extracted data to identify common trends, challenges, and gaps in the existing research. Based on our analysis, we developed a taxonomy that categorizes the studies into different subfields such as E-commerce and retail, Social Media, Entertainment and Media, Healthcare, Finance. This taxonomy helps to highlight the areas where significant progress has been made as well as the areas that require further research and development.

3 Results

The review of recent advancements in recommender systems across various domains reveals several key trends and innovations:

In e-commerce and retail, there's a significant shift towards leveraging large language models and graph neural networks to enhance recommendation accuracy and personalization. For instance the generative explore-exploit approach optimizes recommendations by balancing user preferences with systematic exploration.

Social media recommender systems are evolving to address challenges like echo chambers and misinformation spread. The SoMeR framework excels in creating detailed user profiles by integrating various data types, while A-LLMRec enhances collaborative filtering with LLMs to improve recommendations across diverse scenarios.

In the entertainment and media sector, there's a focus on addressing popularity bias and enhancing real-time recommendations. The SUBER framework uses LLMs to simulate human behavior for training reinforcement learning agents, while the Interest Clock method effectively encodes time-aware user preferences for streaming recommendations.

Healthcare recommender systems are increasingly emphasizing fairness and privacy. The FAIIR system demonstrates high accuracy in identifying youth issues in crisis text conversations, while F2PGNN achieves fairness in federated graph-based recommender systems while preserving user privacy.

In the finance sector, recommender systems are being tailored to address unique challenges such as risk management and regulatory compliance. NFT-MARS addresses the dual nature of NFTs as both artwork and financial assets, while MVECF-balances risk and return in portfolio recommenda-

tions.

3.1 E-Commerce and Retail

The e-commerce and retail sectors are experiencing a transformative shift driven by cutting-edge artificial intelligence technologies. Large language models (LLMs) and generative AI (GAI) are at the forefront of this revolution, reshaping recommender systems with unprecedented capabilities. These advancements are addressing longstanding challenges in personalization, cold-start problems, and real-time optimization. Novel approaches like Meta-path-guided Identifier (META ID) and A-LLMRec are pushing the boundaries of recommendation accuracy and diversity by leveraging out-of-vocabulary tokens and aligning traditional collaborative filtering with LLM token spaces. Meanwhile, frameworks such as the Low-rank Online Assortment with Dual-contexts (LOAD) and generative explore-exploit methods are optimizing real-time recommendations by intelligently balancing user preferences with strategic exploration. As these technologies mature, they promise to deliver more engaging, personalized, and efficient shopping experiences, potentially reshaping the future of digital commerce.

- Huang et al. (2024) introduce Meta-path-guided Identifier (META ID) is a framework for improving large language models for recommendation tasks by constructing user and item IDs using out-of-vocabulary (OOV) tokens. META ID captures user-item correlations and enhances diversity by learning user and item representations from meta-paths, clustering them, and assigning hierarchical OOV tokens. Integrating these OOV tokens into the LLM's vocabulary enables better capture of user-item relationships during fine-tuning. Experiments demonstrate META ID's strong performance across various recommendation tasks.

- This survey examines the current landscape and future directions for integrating generative AI (GAI) into industrial social and e-commerce recommender systems (Recsys). It discusses the challenges posed by the complex infrastructure and product sophistication of modern Recsys, and presents practical solution frameworks based on industry experiences. Key areas covered in this survey (Xu et al., 2024a) include: 1) Enhancing personalized recommendation using GAI for data processing, feature engineering, and modeling; 2) Augmenting Recsys' content curation capabilities through retrieval-augmented generation (RAG); 3) Facilitating interactive recommendation and active feedback loops with autonomous AI agents; 4) Addressing responsible AI and human-AI alignment challenges. The survey also highlights open problems and practical considerations to guide the effective adoption of GAI in real-world Recsys.

- Senel et al. (2024) introduce a training-free approach for optimizing generative recommender systems using large language models and user feedback loops. The proposed generative explore-exploit method iteratively refines the generated item pool by exploiting high-performing items and actively exploring user preferences to improve recommendation quality. Experiments on question generation for e-commerce and general knowledge domains, with user feedback simulated using click-through rate (CTR), demonstrate that the LLM-based explore-exploit approach can effectively adapt recommendations to match hidden user preferences and consistently increase CTR. The results highlight the importance of generative exploration in discovering user preferences, avoiding the pitfalls of greedy exploit-only approaches.

- Sinha and Gujral (2024) introduce PAE, a product attribute extraction framework for future trend reports consisting of text and images in PDF format. PAE addresses the challenges of extracting attributes from complex PDF layouts and mapping them to existing product catalogs. The proposed approach involves four steps: 1) extracting text and images from PDF files, 2) extracting attributes using Large Language Models, 3) consolidating attributes into categories, and 4) matching attributes with the product catalog using BERT embeddings. Experiments on real-world datasets demonstrate that PAE outperforms state-of-the-art methods in attribute extraction accuracy and efficiency. The framework provides valuable insights for retailers to plan future product assortments based on upcoming trends.

Domain	Applications	Key Techniques	Datasets Used	Evaluation Metrics	Limitations
E-commerce and Retail	Price Prediction Personalized Marketing Product Recommendation Customer Segmentation	• Collaborative Filtering • Multi-modal Learning • Graph Neural Networks • Large Language Models • Meta-path-guided Identifier • Retrieval-Augmented Generation • Low-rank Online Assortment Models • Alignment of Embeddings with LLMs	Amazon MovieLens Ciao Expedia Retail datasets Real-world transaction data	Recall@K Precision@K NDCG@K Click-through rate Conversion rate Regret bounds	• Cold Start Problem • Data Sparsity • High Computational Cost • Limited Feedback • Privacy Concerns • Vulnerability to Attacks
Social Media	Sentiment Analysis Fake News Detection Friend Recommendations Content Recommendation User Profiling Trust and Polarization Modeling	• Multi-view User Representation Learning • LLM-enhanced Collaborative Filtering • Diffusion-based Social Denoising • Algorithmic Audits • Game-theoretic User Strategization Models • Graph Neural Networks • Transformers	Yelp Twitter Stocktwits Tumblr TikTok data Facebook data	Accuracy F1 score Precision Recall Network Density Engagement Metrics Regret Bounds	• High Noise Level • Semantic Gap Between Modalities • Echo Chambers • Misinformation Spread • Privacy Issues • Data Poisoning Attacks
Entertainment and Media	Content Recommendation User Engagement Streaming Recommendations Personalization Music/Video Suggestions Popularity Bias Mitigation	• LLM-simulated User Behavior • Interest Clock Method • Sequential Recommendation Models • Popularity Bias Metrics • Transformer-based Models • Collaborative Filtering • Deep Learning	MovieLens 25M Douyin Music App data Netflix data Spotify data LastFM data	MAP@K NDCG@K User Active Days App Duration Engagement Metrics Retention Rate Popularity Bias Measures	• Popularity Bias • Over-personalization • Lack of Real-time Feedback • Cold Start Problem • Scalability Issues • Evaluation Challenges
Health Care	Disease Prediction Patient Monitoring Medication Recommendation Treatment Center Assignment Clinical Decision Support Fairness in Recommendations	• Empirical Soft Regret Loss Function • Domain-adapted Transformer Models • LEADER Framework • Federated Graph-based Recommender Systems • DKINet • Deep Learning • Reinforcement Learning • Ensemble Methods	MIMIC-III MIMIC-IV eICU SEER dataset Electronic Health Records Clinical Trial Data	Accuracy Precision Recall ROC-AUC F1 score Survival Rate Fairness Metrics	• Privacy Concerns • Data Imbalance • High Dimensionality • Fairness Issues • Regulatory Compliance • Interpretability • Lack of Generalizability

Table 1: This table presents the key applications, techniques, datasets, evaluation metrics, and limitations of recommender systems across domains like E-commerce, Social Media, Entertainment, and Healthcare. It highlights advanced methods such as collaborative filtering, large language models, and deep learning, with datasets including MovieLens, Twitter, and MIMIC-III.

Domain	Applications	Key Techniques	Datasets Used	Evaluation Metrics	Limitations
Finance	Stock Price Prediction Portfolio Management NFT Recommendations Financial Product Recommendation Risk Assessment Investment Advice	• Recommender Systems in Financial Trading • NFT-MARS • Mean-Variance Embedding Collaborative Filtering • Knowledge Graph-driven Recommender Systems • Zero-shot Learning Approaches • Time Series Analysis • Sentiment Analysis • Reinforcement Learning	S&P 500 Hong Kong Stock data Zephyr NFT transaction data Financial Statements Market Data	F1 score True Positive Rate Precision Recall Sharpe Ratio Return on Investment Mean Squared Error	• Model Misspecification • Overfitting • Reliance on Historical Data • Market Volatility • Limited User Feedback • Regulatory Constraints • Ethical Considerations
Travel & Tourism	Personalized Itinerary Planning Tour Recommendation Passenger Guidance Transportation Optimization Context-aware Recommendations Route Planning	• Context-aware and Configurable TRSs • BERT-based Models • Knowledge Graph Convolutional Networks • Graph Neural Networks • Deep Reinforcement Learning • Non-stationary Bandits	Flickr Data Tourism Datasets Beijing Subway Network Data Real-world Road Networks Theme Park Data	F1 Score Accuracy Travel Time Reduction User Satisfaction Engagement Metrics Computational Efficiency	• Data Sparsity • Dynamic Environments • Scalability • Real-time Constraints • Personalization Challenges • Evaluation Complexity
Education	Course Recommendation Personalized Learning Paths Gamification in E-learning MOOC Recommendations Knowledge Tagging Fairness in Recommendations	• Integration of LLMs (Tailor-Mind, RAMO) • Reinforcement Learning • Graph Reasoning (UPGPR) • SimCE Loss Function • Collaborative Filtering • Gamification Models • Explainable AI	Coursera Dataset COCO XuetangX MathKnowCT Educational Platforms MOOC Data	F1 Score NDCG Hit Ratio MAE User Engagement Learning Efficiency Fairness Metrics	• Cold Start Problem • Data Privacy • Interpretability • Balancing Accuracy and Fairness • Scalability • Lack of Real-world Deployment
Bandits	Online Testing Recommender Systems Marketing Optimization Budget Allocation in Advertising Dynamic Decision-Making Exploration vs. Exploitation	• Multi-Armed Bandits • Epsilon-Greedy • Upper Confidence Bound • Thompson Sampling • Contextual Bandits • Bandits with Budgets • Primal-Dual Approach • Reinforcement Learning	Amazon Music Recommendation Data Stitch Fix Experimentation Data Facebook Advertising Data Synthetic Data for Simulations	Regret Bounds Conversion Rate Click-through Rate Revenue Maximization User Engagement Metrics Cumulative Reward	• Computational Complexity • Need for Large Data • Scalability Issues • Balancing Exploration and Exploitation • Implementation Complexity • Ethical Considerations

Table 2: This table summarizes key applications, techniques, datasets, evaluation metrics, and limitations of recommender systems in domains such as Finance, Travel & Tourism, Education, and Bandits. Applications include stock price prediction, itinerary planning, and course recommendation. It highlights techniques like collaborative filtering, graph neural networks, and multi-armed bandits.

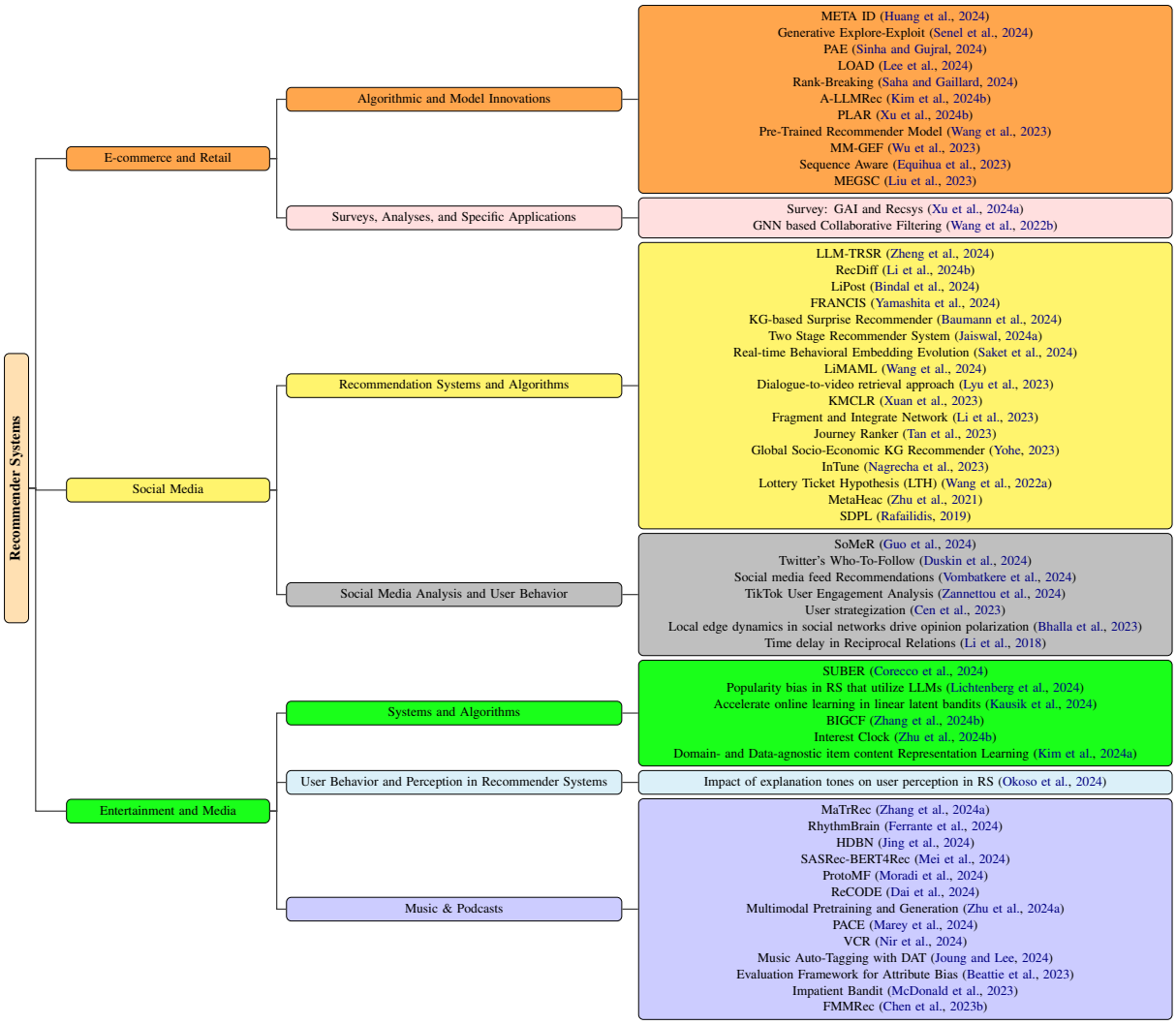

Figure 1: This taxonomy categorizes the cutting edge techniques on recommender systems into different subfields, including E-Commerce and Retail, Social Media, Entertainment and Media. It illustrates the distribution of research focus across various domains and highlights the advancements, challenges, and gaps in each subfield.

- In their paper Lee et al. (2024) propose a new Low-rank Online Assortment with Dual-contexts (LOAD) model to address the challenge of delivering real-time personalized recommendations from vast catalogs in e-commerce. Their model incorporates both user and item features, utilizing a low-rank structure to efficiently handle high-dimensional scenarios. The authors introduce the Explore-Low-rank-Subspace-then-Apply-UCB (ELSA-UCB) algorithm, which estimates intrinsic subspaces and employs an upper confidence bound approach to balance exploration and exploitation in online decision making. Theoretical analysis establishes a regret bound of $\mathcal{O}((d_1+d_2)r\sqrt{T})$, where d_1 and d_2 represent the dimensions of user and item features, r is the rank of the parameter matrix, and T is the time horizon. This bound significantly improves upon prior literature by leveraging the low-rank structure. Extensive simulations and an application to the Expedia hotel recommendation dataset demonstrate the advantages of the proposed method.

- Saha and Gaillard (2024) addresses the active online assortment optimization problem with preference feedback, aiming to develop practical, efficient, and optimal algorithms that overcome limitations of existing methods. The authors propose a novel approach that utilizes the concept of "Rank-Breaking" to establish tight concentration guarantees for estimating score parameters of the Plackett-Luce (PL) model. The proposed algorithms, AOA-RB-PL and AOA-RB-PL-Adaptive, are designed to be practical, provably optimal (up

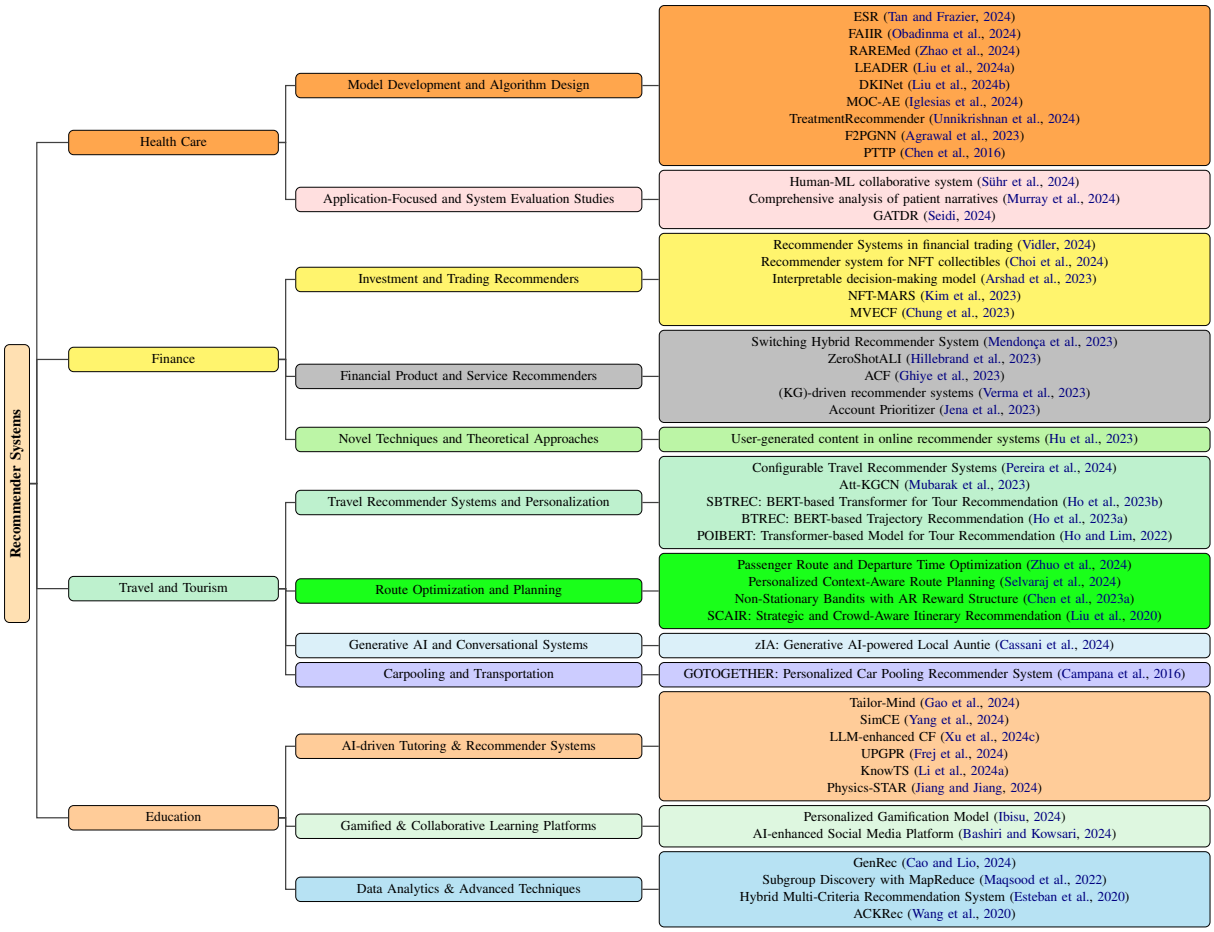

Figure 2: This taxonomy categorizes the cutting edge techniques on recommender systems into different subfields, including Finance, Travel & Tourism and Education, It illustrates the distribution of research focus across various domains and highlights the advancements, challenges, and gaps in each subfield.

to logarithmic factors), and do not require the restrictive assumption of a strong default "No-Choice" item. The algorithms employ an optimistic approach based on Upper Confidence Bound (UCB) estimates of PL parameters and avoid repeating the same subsets multiple times. Extensive empirical evaluations demonstrate the superior performance of the proposed methods compared to existing baselines, highlighting their effectiveness in real-world assortment optimization scenarios.

- A-LLMRec, an efficient all-round Large Language Model based recommender system that excels in both cold and warm scenarios. (Kim et al., 2024b) A-LLMRec aligns the user/item embeddings of a pre-trained collaborative filtering recommender system (CF-RecSys) with the token space of an LLM, enabling the LLM to leverage the collaborative knowledge for recommendation. The alignment network is the only trainable component, making A-LLMRec model-agnostic and efficient. Experiments on various real-world datasets demonstrate A-LLMRec's superiority in cold/warm, few-shot, cold user, and cross-domain scenarios. A-LLMRec also generates natural language outputs based on its understanding of collaborative knowledge. The proposed approach outperforms state-of-the-art CF-RecSys, modality-aware, and LLM-based recommenders while being 2.53 times faster in training and 1.71 times faster in inference compared to fine-tuning LLMs.

- Xu et al. (2024b) explore the synergies between large language models and collaborative filtering algorithms for enhancing e-commerce recommendation systems. The authors propose a framework, PALR, that combines user behavior data with LLMs to generate personalized item recommendations. LLMs offer advantages such as perceptual learning, easy integration of multimodal sig-

nals, knowledge transfer for cold-start scenarios, and natural language explanations. The paper reviews LLM techniques for collaborative filtering from pre-training, fine-tuning, and prompting paradigms. Experiments compare traditional user-based and item-based collaborative filtering algorithms with their LLM-enhanced counterparts, demonstrating improved accuracy and recommendation performance. The integration of LLMs with collaborative filtering holds promise for delivering more accurate and personalized recommendations, ultimately driving sales and user satisfaction in e-commerce platforms.

- Jaiswal (2024b) establishes the asymptotic convergence properties of two-tower recommender systems, which employ deep neural networks to learn low-dimensional representations of users and items for collaborative filtering. The authors quantify the approximation and estimation errors of the model, showing that its convergence rate depends on the smoothness of the optimal recommender system and the intrinsic dimensionality of user and item features. Under certain conditions, the convergence rate can be as fast as $O(|\Omega|^{-1}(\log|\Omega|)^2)$, where Omega is the set of observed ratings. The paper provides statistical guarantees for the two-tower model, justifying its effectiveness in real-world applications. Experiments on synthetic and real-world datasets demonstrate the model's superior performance compared to existing methods, especially in scenarios with data scarcity and cold-start users.

- Wang et al. (2023) propose a novel pre-trained recommender model framework that can generalize to unseen users and items within a dataset as well as across different datasets, without relying on any auxiliary user or item information. The key insight is to leverage universal statistical characteristics of the user-item interaction matrix, such as user and item activity distributions and co-occurrence patterns, to learn transferable representations. The authors introduce three types of dataset-agnostic features: 1) activity-based features derived from user and item marginal activity distributions, 2) co-occurrence based features capturing joint interaction patterns, and 3) interaction-based edge representations. A simple pre-trained model consisting of feed-forward layers is trained on these features. Extensive experiments on five real-world datasets demonstrate strong zero-shot generalization performance within and across domains, as well as the ability to enhance state-of-the-art neural collaborative filtering models in the traditional full-information setting. This work pioneers the development of universal recommender systems that can be adapted to new applications with minimal or no retraining.

- Wu et al. (2023) propose MM-GEF, a novel graph-based method for multi-modal recommendation that effectively incorporates cross-modal relations and implicit collaborative signals into an early-fused item graph structure. MM-GEF leverages pre-trained contrastive visual-language representations (CLIP) to guide the multi-modal feature extraction and fusion process. The fused multi-modal features are then compressed into low-dimensional item representations. MM-GEF builds two homogeneous item graphs based on the fused multi-modal features and collaborative filtering signals respectively, and combines them using a soft attention mechanism. Graph convolution is applied on the unified item graph to inject high-order item relationships into the item representations. Finally, the refined item embeddings are integrated into a collaborative filtering model for recommendation. Extensive experiments on four datasets demonstrate the superiority of MM-GEF over state-of-the-art multi-modal recommender systems, highlighting the benefits of early multi-modal fusion and the incorporation of collaborative signals into the item graph structure learning process.

- Equihua et al. (2023) present a novel sequence-aware recommender system that models user-item interactions over time using recurrent neural networks and natural language processing techniques. Unlike traditional collaborative filtering and matrix factorization approaches, this method processes customer transactions as sequential information, considering each item in the product catalog as a token. Sequences of tokens are generated

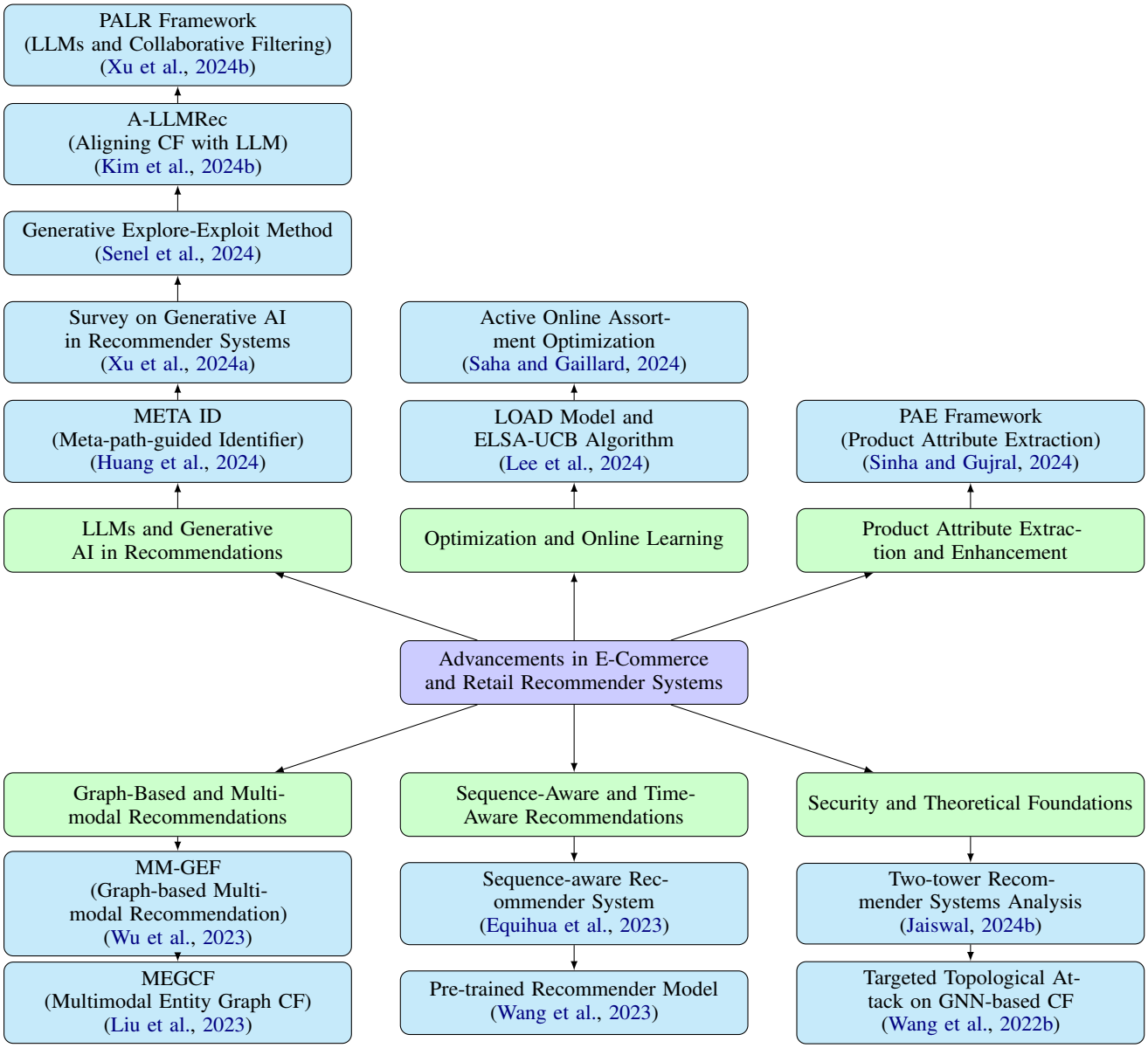

Figure 3: Concept map illustrating the advancements in E-Commerce and Retail recommender systems. The map is divided into six primary categories: **LLMs and Generative AI in Recommendations**, **Optimization and Online Learning**, **Product Attribute Extraction and Enhancement**, **Graph-Based and Multimodal Recommendations**, **Sequence-Aware and Time-Aware Recommendations**, and **Security and Theoretical Foundations**. Under each category, key research studies and methods are outlined with citations, showing the progression and relationships between different approaches. Cross-connections indicate how certain methodologies influence or relate to others across categories, highlighting the interdisciplinary nature of recent advancements in e-commerce recommender systems.

for each user based on their transaction history. The proposed model, consisting of an embedding layer, LSTM units, and a feed-forward network, is trained to predict the probability of future user-item interactions. Extensive experiments on two real-world retail datasets and the MovieLens dataset demonstrate that this approach maintains predictive performance comparable to other recommendation techniques, while being particularly effective for the top item recommendations. A live A/B test with a large UK retailer shows a significant increase in revenue and precision compared to a matrix factorization-based system. The sequence-aware recommender system is more suitable for settings where user preferences change rapidly, users interact repeatedly with specific items, and providing relevant recommendations is crucial for improving customer engagement and sales.

- Liu et al. (2023) propose MEGCF, a novel graph-based method for multimodal recommendation that addresses the mismatch problem between multimodal feature extraction (MFE) and user interest modeling (UIM). MEGCF extracts semantic-rich entities from multimodal data and integrates them into the user-item interaction graph to capture multimodal semantic correlations. A symmetric linear Graph Convolution Network (GCN) module is constructed to perform message propagation over the graph, capturing high-order semantic correlations and collaborative filtering signals. Sentiment information from reviews is used to weight neighbor aggregation in the GCN, reflecting the overall quality of items. Extensive experiments on three datasets demonstrate the effectiveness of MEGCF, outperforming state-of-the-art multimodal and graph-based collaborative filtering methods. This work highlights the importance of solving the mismatch problem between MFE and UIM, and the benefits of incorporating semantic entities and sentiment information for multimodal recommendation.

- Wang et al. (2022b) revisit the item promotion task for GNN-based collaborative filtering recommender systems from a novel targeted topological attack perspective. The authors formulate the task as a constrained optimization problem to maximize a target item's popularity by manipulating a limited number of user-item interactions in the graph topology. They employ gradient-based optimization to find a solution and propose a node masking mechanism to enhance the attack ability by resolving noisy gradient effects. A resource-efficient approach is also designed to enable scalability for large-scale collaborative filtering systems. Extensive experiments on two real-world datasets demonstrate the effectiveness of the proposed attack in significantly promoting target items' popularity, outperforming baseline methods. The findings highlight the vulnerability of GNN-based collaborative filtering models to this new type of targeted topological attack, raising concerns about their security and adversarial robustness in practical deployment scenarios.

3.2 Social Media

The landscape of social media and recommendation systems is undergoing a profound transformation, driven by cutting-edge artificial intelligence and machine learning technologies. Recent research has focused on developing more sophisticated user representation models, leveraging large language models for enhanced recommendations, and addressing critical issues such as trust, echo chambers, and polarization. Novel frameworks like SoMeR and A-LLMRec are pushing the boundaries of user modeling and recommendation accuracy by incorporating multi-modal data and aligning collaborative filtering with large language models. Concurrently, researchers are tackling challenges related to data sparsity, content understanding, and the ethical implications of AI-driven social media ecosystems. Studies on platforms like Twitter and TikTok are shedding light on the complex interplay between recommendation algorithms, user behavior, and social dynamics. As these technologies evolve, they're not only enhancing user experiences but also raising important questions about privacy, algorithmic bias, and the potential for manipulation in digital spaces. This rapidly advancing field is reshaping our understanding of online interactions and paving the way for more personalized, engaging, and potentially more responsible social media platforms.

- Guo et al. (2024) present SoMeR, it is a multi-view user representation learning framework

for social media that incorporates temporal activities, text content, profile information, and network interactions to learn comprehensive user portraits. It encodes user post streams as sequences of timestamped textual features, uses transformers to embed this along with profile data, and jointly trains with link prediction and contrastive learning objectives to capture user similarity. The framework's versatility is demonstrated through two applications: 1) Identifying inauthentic accounts involved in coordinated influence operations, and 2) Measuring increased polarization in online discussions after major events. SoMeR's ability to holistically model users enables new solutions to important problems around disinformation, societal tensions, and online behavior understanding.

- Duskin et al. (2024) present an algorithmic audit of Twitter's Who-To-Follow friend recommendation system, investigating its impact on political echo chambers and exposure to false and misleading content during the 2022 U.S. midterm elections. The authors created automated accounts that initially followed left and right affiliated politicians, then grew their networks using either the platform's recommendation algorithm or social endorsement. They found that while the recommendation algorithm leads to dense and reciprocal network structures resembling echo chambers, it results in less political homogeneity and fewer opportunities to encounter false or misleading election narratives compared to social endorsement-based growth. The findings suggest that the recommendation algorithm alone is not the key driver of observed political echo chambers on Twitter.

- Zheng et al. (2024) present LLM-TRSR is a novel framework for harnessing large language models for text-rich sequential recommendation that addresses the challenges of over-length limitations, extensive time and space overheads, and suboptimal model performance. It segments user behavioral history, employs an LLM-based summarizer to summarize user preference using hierarchical or recurrent summarization techniques, and constructs a prompt encompassing the user preference summary, recent interactions, and candidate item information for an LLM-based recommender fine-tuned using supervised fine-tuning with LoRA for parameter-efficient fine-tuning. Experiments on two public datasets demonstrate the effectiveness of the approach in capturing both long-term and short-term user preferences for improved text-rich sequential recommendations.

- Vombatkere et al. (2024) propose a general framework to examine social media feed recommendations for a user as a timeline, labeling items as the result of exploration vs. exploitation of the user's interests and introducing metrics to capture the extent of personalization. Applying the framework to real TikTok user data and validating the results using automated TikTok bots and a randomized baseline, the authors find that TikTok's recommendation algorithm exploits users' interests in 30-50% of recommended videos in the first thousand, with liking and following being the primary drivers of personalization. The proposed framework can be used to audit and understand personalization in social media feeds, aiding transparency efforts, providing user insights, and enabling algorithmic audits by policymakers and researchers.

- Li et al. (2024b) propose RecDiff, is a novel diffusion-based social denoising framework for recommendation that utilizes a hidden-space diffusion paradigm to alleviate the noisy effect in the compressed and dense representation space. By performing multi-step noise diffusion and removal, RecDiff identifies and eliminates noise from encoded user representations, even when noise levels vary. The diffusion module is optimized in a downstream task-aware manner, maximizing its ability to enhance the recommendation process. Extensive experiments demonstrate RecDiff's superiority in recommendation accuracy, training efficiency, and denoising effectiveness compared to state-of-the-art baselines.

- Bindal et al. (2024) introduce LiPost, a content understanding model developed by LinkedIn to enhance semantic comprehension in recommendation systems. The model employs multi-task contrastive learning, fine-tuning a pre-trained multilingual BERT on diverse semantic labeling tasks including

interest classification, storyline categorization, hashtag prediction, and search relevance. LiPost demonstrates improved performance across all individual tasks compared to single-task models, showcasing positive transfer learning. The model exhibits strong zero-shot learning capabilities and enhanced multilingual support, outperforming the baseline on 50 languages with a 9.6% relative improvement. LiPost achieves comparable performance to state-of-the-art OpenAI embeddings (e.g., ADA_002) on LinkedIn-specific tasks while offering 30x compression with only 50-dimensional embeddings, crucial for meeting production latency requirements. The model's deployment in LinkedIn's feed ranking system resulted in a 0.1% increase in user sessions and a 0.21% increase in professional interactions. The authors discuss ethical considerations, including fairness evaluations across language groups, and acknowledge limitations such as the current lack of multimodal capabilities and a limited context window. Future work aims to incorporate multimedia content, explore new architectures, and implement online triplet mining for continued improvement.

- Yamashita et al. (2024) introduce FRANCIS, a novel data poisoning attack framework targeting online job platforms, demonstrating critical vulnerabilities in career prediction models. The authors propose three attack scenarios: company promotion, company demotion, and user promotion attacks, exploiting the ease of creating fake accounts and the unrestricted nature of resume content. FRANCIS employs a probabilistic job trajectory generator, a reality regulation function, and an attack module to create realistic fake resumes that manipulate prediction outcomes. Experiments conducted on real-world datasets from tech and business sectors show that FRANCIS outperforms baseline methods, achieving improvement rates of up to 23.17 at 10% injection, 4.98 at 1%, and 1.32 at 0.1% injection for company promotion attacks. The framework's effectiveness persists across various career prediction models (NEMO, AHEAD, NAOMI) and injection rates, with minimal impact on overall model performance, making detection challenging. The study reveals that even small-scale attacks (0.01% injection) can significantly alter prediction results, highlighting the urgent need for enhanced security measures in online job platforms. The authors discuss ethical considerations, limitations, and future research directions, emphasizing the broader implications for HR-related tasks and the importance of developing robust defense mechanisms.

- (Baumann et al., 2024) propose a novel approach to introduce a configurable degree of surprise into recommender systems using knowledge graphs (KGs) and complex network metrics. The authors construct KGs from user profiles and item catalogs, then rerank recommendations based on their impact on structural graph metrics when added to user profile subgraphs. They evaluate their method on LastFM and Netflix datasets, focusing on metrics like betweenness centrality, node/edge counts, and degree-based features. The study finds that ranking items by ascending betweenness centrality leads to more diverse and unexpected recommendations while maintaining low correlation with standard recommender outputs. Experiments show improvements in unexpectedness and intra-list diversity for LastFM data, and significant changes in nDCG scores for both datasets, indicating the method's ability to introduce surprise. The approach is modular and can be applied on top of any existing recommender system. The authors discuss computational considerations and suggest node/edge counts or degree-based features as efficient alternatives to betweenness centrality for large-scale applications. Overall, the study demonstrates that network-level metrics in KGs can effectively influence the degree of surprise in recommendations, with betweenness centrality showing the strongest effect.

- Jaiswal (2024a) present a theoretical analysis of two-stage recommender systems, focusing on their asymptotic behaviors and convergence properties. The authors introduce a model called Journey Ranker, which uses two deep neural networks to embed users and items into a low-dimensional space. They prove that this model converges strongly to an optimal recommender system, with the convergence rate dependent on the smooth-

ness of the true model and the intrinsic dimensions of user and item features. The paper establishes upper bounds on approximation and estimation errors, showing that the convergence rate can reach $O(|\Omega|^{-1}(\log|\Omega|)^2)$ as smoothness approaches infinity, where Ω is the set of observed ratings. The authors demonstrate that finite depths of the neural networks are sufficient for approximating the true model, while the widths increase at a rate of $O(|\Omega|^{d_{ui}/(2\beta+d_{ui})}\log|\Omega|)$, where d_{ui} is the maximum intrinsic dimension and β is the smoothness parameter. Experimental results on synthetic and real-world datasets show that Journey Ranker outperforms several baseline models, particularly in sparse rating scenarios and when addressing the cold-start problem. The paper provides a rigorous theoretical foundation for understanding the behavior of two-stage recommender systems and offers insights into their practical implementation and performance.

- Saket et al. (2024) investigate the evolution of behavioral embeddings in short-video recommendation systems, comparing batch and real-time update strategies using data from ShareChat, a platform with over 180 million users. The study focuses on three key aspects: embedding maturity, trajectory of embeddings over time, and the distribution of ℓ_2-norms. The authors find that real-time embeddings mature significantly faster, requiring only 20% of the user interactions needed by batch learning. They identify that the highest information updates occur earlier in real-time learning, leading to quicker convergence. The analysis of ℓ_2-norm distributions reveals that batch-trained embeddings suffer from higher norms for high-impression posts, contributing to popularity bias. The findings are validated using user engagement metrics, including video click rates and successful video play percentages. The study demonstrates that real-time updates lead to better performance in lower view buckets, resulting in improved content targeting and higher user engagement. The authors also discuss the implications of these findings on popularity bias and provide insights into designing effective recommendation systems for short-video applications. The paper concludes by suggesting future research directions, including determining the minimum number of views required to understand content complexity, exploring techniques to reduce update frequency, and investigating optimal strategies for content exposure to maximize learning outcomes.

- Zannettou et al. (2024) present the first large-scale empirical study of user engagement with short-format videos on TikTok using data donated by 347 real users, comprising 9.2M video views. The researchers developed a data donation system to collect and analyze user behavioral traces, focusing on metrics like time spent, video views, attention (videos watched to completion), and interaction (likes). Key findings include: users' daily average time spent and number of videos viewed increased over time (2x increase after 80 days); attention remained stable at around 45% of videos watched to completion; users watched more videos to completion from non-following accounts (44-46%) compared to following accounts (38-42%); and user interaction through likes increased over time (2x for following accounts, 1.5x for non-following accounts after 120 days). The study also found that only 10.3% of video views were from accounts users followed, despite users following more accounts over time. Videos from non-following accounts were significantly more popular platform-wide. The researchers discuss implications for user well-being, algorithmic design, and the potential for addictive behavior on short-format video platforms. They also share insights on designing and implementing data donation systems for social media research.

- Wang et al. (2024) introduce LiMAML, a novel meta-learning approach for personalizing deep recommender systems at scale. LiMAML extends Model-Agnostic Meta-Learning (MAML) by dividing the network into a meta block and a global block, enabling efficient personalization for millions of users or entities. The meta block is meta-learned and fine-tuned offline for each task (e.g., user), producing fixed-size meta embeddings that are used as inputs to the global block during online inference. This approach overcomes the storage and latency challenges of deploy-

ing MAML in large-scale production systems. The authors demonstrate LiMAML's effectiveness on multiple LinkedIn applications, including push notification and in-app recommendation tasks, showing consistent improvements over strong baselines in both offline and online A/B tests. Notably, LiMAML achieves significant performance gains for infrequent users and new members, demonstrating its ability to personalize models with limited data. The paper also details production insights, training speed-up techniques, and ablation studies on hyperparameters and model components. Overall, LiMAML presents a scalable and effective framework for personalizing deep recommender models in industrial settings, addressing the critical need for user-specific and frequently updated models in large-scale recommendation systems.

- Cen et al. (2023) study the implications of user strategization when interacting with data-driven platforms through a game-theoretic model. The authors find that while strategization can benefit platforms in the short term by providing cues the platform would otherwise miss, it ultimately corrupts the platform's data, hurting its ability to make counterfactual decisions. They connect this phenomenon to user trust and show that designing trustworthy algorithms, formalized as not incentivizing strategization and ensuring a minimum user utility, can mitigate the effects of strategization. The paper provides recommendations for trustworthy design, such as offering multiple algorithms and feedback mechanisms, to induce approximately exogenous user behavior while uncovering what users find untrustworthy.

- Lyu et al. (2023) introduce a novel dialogue-to-video retrieval approach that incorporates structured conversational information from user-generated dialogues to improve video retrieval performance. The proposed model sequentially encodes each turn of the dialogue to obtain a dialogue-aware query representation and calculates its similarity with individual video frames to obtain a weighted video representation. Experiments on the AVSD dataset demonstrate that using dialogues as search queries significantly improves retrieval performance compared to using plain-text queries. The model achieves state-of-the-art results, outperforming previous approaches by 0.7%, 3.6%, and 6.0% on R@1, R@5, and R@10 metrics, respectively. The study highlights the importance of utilizing dialogue information in video retrieval systems, especially for recommendation purposes.

- Xuan et al. (2023) propose KMCLR, a novel multi-behavior recommendation framework that integrates contrastive learning with knowledge graph information to alleviate data sparsity issues and capture user preferences from various perspectives. The model consists of a multi-behavior learning module that extracts personalized behavior information, a knowledge enhancement module that derives robust knowledge-aware item representations, and a joint learning module that optimizes the model using two contrastive learning tasks and a loss paradigm. Experiments on three real-world datasets demonstrate that KMCLR outperforms state-of-the-art methods by modeling the commonalities between behaviors and differences between users. The study also shows that KMCLR effectively alleviates data sparsity and noise issues caused by the introduction of auxiliary information.

- Bhalla et al. (2023) introduce a model that explores how local edge dynamics in social networks drive opinion polarization. Building on the Friedkin-Johnsen opinion model, the authors integrate mechanisms like confirmation bias and friend-of-friend link recommendations. Simulations on synthetic and real-world graphs demonstrate that the combination of these dynamics significantly increases polarization by forming distinct clusters of similar opinions while diminishing inter-group connectivity. The model's realism is validated through comparisons with real social network structures, showing shifts towards natural degree distributions and clustering coefficients. This study provides insights into the underlying mechanisms of opinion polarization driven by both human behavior and platform algorithms.

- Li et al. (2023) introduce Fragment and Integrate Network (FIN), a novel spatial-temporal modeling approach for click-through rate

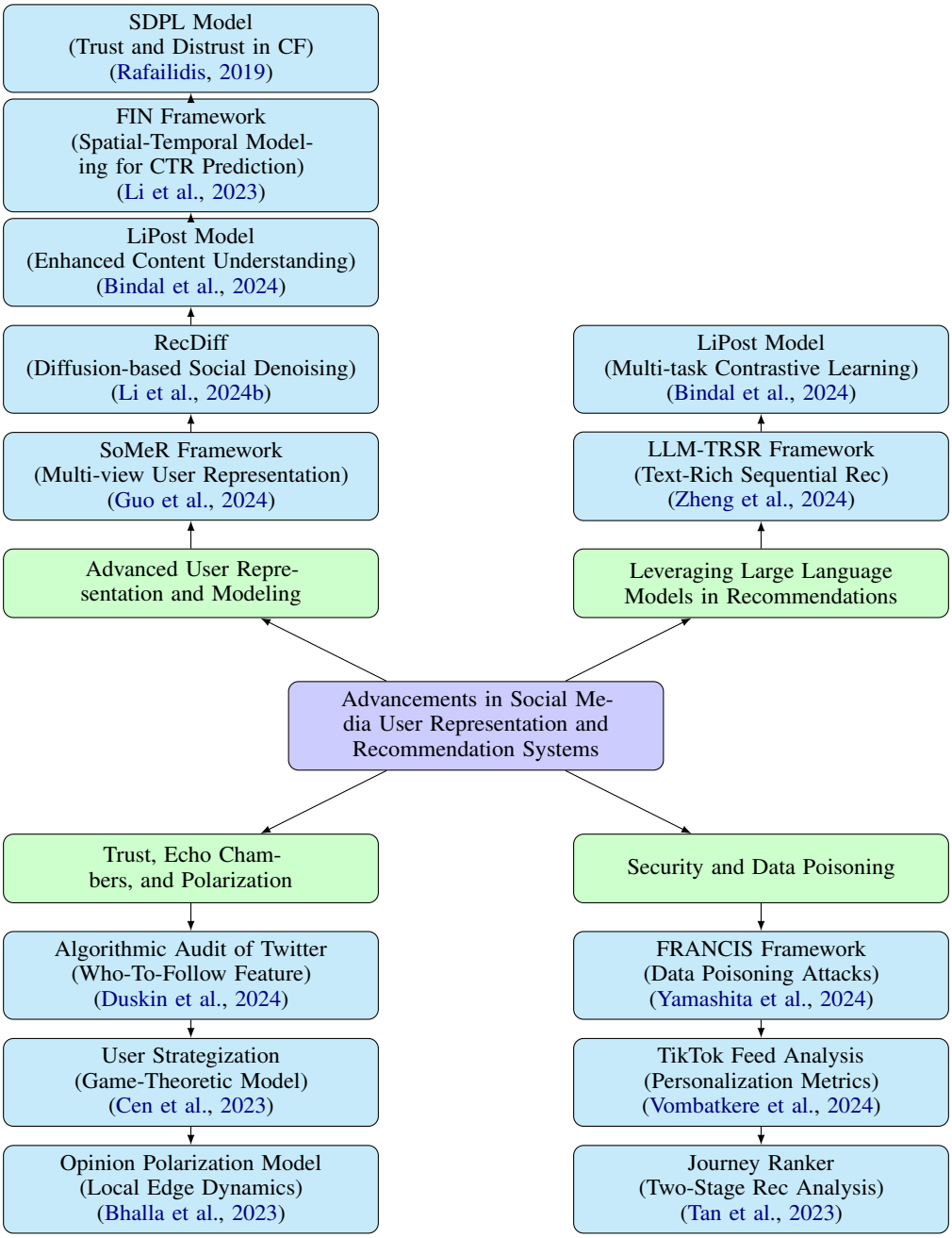

Figure 4: Concept map illustrating the advancements in social media user representation and recommendation systems. The map is divided into key areas: **Advanced User Representation and Modeling**, **Leveraging Large Language Models in Recommendations**, **Trust, Echo Chambers, and Polarization**, and **Security and Data Poisoning**. Under each category, significant research studies and methods are outlined, showing the progression and relationships between different approaches. Cross-connections indicate how certain methodologies influence or relate to others across categories, highlighting the interdisciplinary nature of recent advancements in social media recommender systems.

(CTR) prediction in online food ordering services. FIN addresses the challenge of efficiently modeling long sequential behavior data with rich spatial-temporal information. The architecture consists of two main components: (1) Fragment Network (FN), which extracts Multiple Sub-Sequences (MSS) from lifelong sequential behavior data and captures specific spatial-temporal representations using simplified and multi-head attention mechanisms; and (2) Integrate Network (IN), which builds an integrated sequence utilizing spatial-temporal interactions on MSS to capture comprehensive spatial-temporal representations. FIN incorporates geohash-block, meal-time, short-term, and long-term de-duplicated subsequences to model user preferences effectively. The authors evaluate FIN on two public datasets (Amazon and Google Local) and an industrial dataset from Ele.me, demonstrating superior performance compared to state-of-the-art models like DIN, SIM, ETA, and StEN. In online A/B testing on Ele.me's recommendation advertising system, FIN achieved a 5.7% improvement in CTR and a 7.3% increase in Revenue Per Mille (RPM). The paper also provides insights into practical implementation considerations for large-scale deployment, including real-time prediction optimizations and CUDA Graph optimization for efficient multi-head attention calculations.

- (Tan et al., 2023) introduce a novel multi-task deep learning model architecture for optimizing search ranking in Airbnb's long and exploratory user journey. The model addresses unique challenges in balancing guest and host preferences while guiding users through multiple stages of decision-making. Journey Ranker consists of four modules: (1) a Shared Representation Module for feature embedding, (2) a Base Module that models positive milestones using a chain of conditional probabilities, (3) a Twiddler Module that handles negative milestones, and (4) a Combination Module that balances outputs from the Base and Twiddler modules using context-dependent coefficients. The architecture leverages intermediate guest actions as both positive and negative milestones to better progress users towards successful bookings. Offline experiments show Journey Ranker achieves a 0.48% improvement in NDCG over the baseline with only a 9.2% increase in parameters. Online A/B tests demonstrate significant gains across multiple Airbnb products: 0.61% increase in uncancelled bookings for stays, 2.0% for experiences, and 9.0% for online experiences. The model's modular and extensible design allows easy application to various use cases beyond search ranking, including email marketing. The authors provide detailed ablation studies and interpretability analyses, revealing insights into the model's learned behaviors across different stages of the user journey.

- Yohe (2023) propose a novel approach to enhance recommender systems by incorporating global, socio-economic, and cultural factors, particularly for online food ordering platforms. The authors develop a Knowledge Graph (KG) based recommender system that encodes user interactions and item catalogs, exploring how network-level metrics on KGs can influence the degree of surprise in recommendations. They hypothesize that surprisingness correlates with certain network metrics, treating user profiles as subgraphs within a larger catalog KG. The study introduces a method to rerank recommendations based on their impact on structural graph metrics, aiming to optimize recommendations to reflect these metrics. Experiments conducted on LastFM listening histories and synthetic Netflix viewing profiles demonstrate that reranking items based on complex network metrics leads to more unexpected and surprising recommendation lists. The authors find that betweenness centrality, when used to rerank recommendations in ascending order, results in the highest levels of unexpectedness and diversity. The research contributes to the field by offering a novel layer on top of existing recommender systems that can incorporate relational information and suggest items with a user-defined degree of surprise. The approach shows promise in addressing the overspecialization issue in recommender systems and enhancing serendipity by uncovering hidden correlations among items.

- Nagrecha et al. (2023) introduce InTune, a novel reinforcement learning-based system

for optimizing data ingestion pipelines in deep learning recommender models (DLRMs). The authors identify that DLRM training is often bottlenecked by online data processing rather than model execution, a problem exacerbated by the unique characteristics of recommendation datasets: scale, reusability, and volatility. Through analysis of real-world DLRM training jobs, they demonstrate that existing tools like AUTOTUNE underperform, particularly with user-defined functions (UDFs) and dynamic resource allocation. InTune addresses these issues by employing a reinforcement learning agent to dynamically distribute CPU resources across pipeline stages. The system uses a carefully designed environment, a lightweight neural network agent, and a reshaped action space to efficiently optimize pipeline throughput. Experimental results show that InTune outperforms state-of-the-art baselines by 1.18-2.29x on real-world workloads, converging on optimized configurations within minutes. It also demonstrates improved scalability, adaptability to resource changes, and robustness against out-of-memory errors. The authors validate InTune's effectiveness across various pipeline complexities, machine sizes, and batch sizes, highlighting its potential to significantly reduce DLRM training costs and times without requiring changes to existing cluster architectures or user workflows.

- Wang et al. (2022a) explore the Lottery Ticket Hypothesis (LTH) in media recommender systems to find winning tickets—sparse subnetworks that can match the performance of the original dense network with fewer parameters. Focusing on the user-item embedding tables of Matrix Factorization (MF) and Light Graph Convolution Networks (LightGCN), the authors apply the Iterative Magnitude Pruning (IMP) algorithm to identify winning tickets. Extensive experiments on three real-world datasets (Yelp2018, TikTok, and Kwai) demonstrate that winning tickets exist widely in recommender models, and the IMP algorithm can reliably find them. The discovered winning tickets significantly outperform the original dense models in terms of memory usage, inference time, and test performance, achieving comparable results with only 3% to 48% of the original parameters. This work is the first to study LTH in recommendation, offering a promising approach to reduce the cost of deploying large-scale media recommender systems.

- Zhu et al. (2021) propose MetaHeac, a novel two-stage framework for audience expansion in recommender systems and advertising platforms. In the offline stage, MetaHeac trains a general model that captures relationships among various marketing campaign tasks using a meta-learning approach with existing campaigns. In the online stage, for a new campaign, a customized model is fine-tuned from the general model using the given seed users to find potential audiences likely to convert. MetaHeac employs a hybrid structure with multiple experts and critics to learn transferable knowledge across tasks. Offline experiments, online A/B testing, and deployment in WeChat demonstrate MetaHeac's superior performance compared to state-of-the-art lookalike methods for content recommendation and advertising.

- Rafailidis (2019) propose a Social Deep Pairwise Learning (SDPL) model for collaborative filtering that leverages both trust and distrust relationships between users to improve recommendation accuracy, especially in scenarios with data scarcity and cold-start users. The model uses a deep learning architecture with a pairwise ranking loss function incorporating multiple ranking criteria based on the preferences of users and their friends and foes. A social negative sampling strategy is employed during training to efficiently learn the model parameters. Experiments on the Epinions dataset demonstrate that SDPL significantly outperforms state-of-the-art methods, with an average improvement of 10.96% over the best baseline. The deep learning architecture and social negative sampling are shown to be key factors in the model's strong performance.

- Li et al. (2018) present the first comprehensive study on predicting the time delay in the formation of reciprocal relations in a directed social network. Using a large-scale dynamic network from Tumblr, the authors uncover several temporal and structural patterns that

influence the delay. They find that the delay is affected by when users join the network, exhibits weekly patterns, and is related to the network structure such as node indegree, outdegree, and number of common neighbors. Based on these findings, they propose a novel model called DPRR (Delay Prediction in Reciprocal Relations) that captures the common and personalized patterns of reciprocal behavior to predict the delay. Experiments demonstrate that DPRR significantly outperforms several baseline methods in accurately predicting the delay in reciprocal link formation.

3.3 Entertainment and Media

The entertainment and media industry is witnessing a transformative era in recommender systems, driven by innovative frameworks and methodologies that leverage cutting-edge artificial intelligence techniques. Large language models (LLMs) are being harnessed in novel ways, such as SUBER's approach to simulating human behavior for training reinforcement learning agents, addressing the challenge of limited real user data. Researchers are tackling longstanding issues like popularity bias, with new metrics like log popularity difference revealing inherent advantages of LLM-based recommenders. Advanced techniques like SOLD are improving online learning by efficiently utilizing offline data, while frameworks such as BIGCF are enhancing user-item interaction modeling by balancing collective and individual factors. In the music and podcast domain, models like MaTrRec are pushing the boundaries of sequential recommendations by combining Mamba and Transformer architectures. Groundbreaking research is also exploring the decoding of music from fMRI data, potentially revolutionizing personalized music experiences. These advancements are not only enhancing recommendation accuracy and user engagement but also addressing critical issues such as fairness, cultural inclusivity, and the ethical implications of AI-driven content curation in the entertainment landscape.

- Corecco et al. (2024) propose SUBER, a novel framework for training and evaluating reinforcement learning-based recommender systems using large language model to simulate human behavior. The LLM-based environment generates synthetic users and item ratings, allowing the training of RL agents without real user data. Extensive ablation studies examine the impact of different framework components, LLM configurations, and fine-tuned user specifications. Experiments on movie and book recommendation tasks, along with human evaluation, demonstrate the effectiveness of using various LLMs to replicate human choices for item recommendations in this synthetic environment. SUBER provides a versatile playground for researchers to improve RL strategies for recommender systems.

- Lichtenberg et al. (2024) examine popularity bias in recommender systems that utilize large language models (LLMs). The authors introduce a principled framework for measuring popularity bias, assessing existing metrics against desiderata of interpretability and statistical robustness and proposing a new metric, log popularity difference, that satisfies these criteria. Comparing a simple LLM-based recommender to traditional models on a movie recommendation task, they find the LLM recommender exhibits less popularity bias even without explicit mitigation. Bias can be further reduced via prompt tuning, presenting a novel opportunity to address the issue. The study highlights the potential of LLMs in recommender systems and the importance of carefully quantifying popularity bias.

- Kausik et al. (2024) tackle the challenge of leveraging offline data to accelerate online learning in linear latent bandits, where unobserved low-dimensional latent states impact reward distributions. The authors prove a de Finetti theorem showing every exchangeable and coherent stateless decision process is a latent bandit. They propose SOLD, a method to estimate the latent subspace spanned by reward parameters from offline data. Two online algorithms, LOCAL-UCB and the more practical ProBALL-UCB, leverage the learned subspace to sharpen optimism and enjoy improved regret guarantees that depend on subspace estimation quality. Experiments on synthetic and real movie recommendation data demonstrate the effectiveness of the approach.

- Zhang et al. (2024b) propose BIGCF, a novel graph collaborative filtering framework that models the individuality and collectivity of intents behind user-item interactions. BIGCF decomposes interaction motivations into col-

lective and individual factors, using collective intents to represent bandwagon and popularity effects while individual intents reflect users' unique preferences and items' specific characteristics. A Gaussian-based graph generation strategy encodes feature distributions to counter data sparsity. Probabilistic graph reconstruction is guided by both individual and collective intents. Graph contrastive regularization in interaction and intent spaces uniformly aligns embeddings without augmentations. Experiments on three real-world datasets demonstrate BIGCF's superior performance compared to state-of-the-art methods.

- Okoso et al. (2024) examine how different tones (e.g., formal, humorous, romantic) in explanations affect user perception in recommender systems across movies, hotels, and home products. Utilizing a large language model to generate fictional items and explanations, the research conducts an online user study to analyze perceived effects, domain differences, and user attributes. Key findings indicate that tone significantly impacts user metrics like transparency, trust, and persuasiveness, with varying effects across different domains. The study also highlights that user attributes, such as age and personality, influence the perceived effectiveness of tonal variations. These insights suggest that tailoring explanation tones to domains and user profiles can enhance the user experience in recommender systems.

- Zhu et al. (2024b) propose Interest Clock, an effective method for time perception in real-time streaming recommendation systems. Interest Clock encodes users' time-aware preferences into hour-level personalized features and uses Gaussian distributions to smooth and aggregate them into a final interest clock embedding based on the current time. This approach transforms time modeling into time-aware feature modeling, addressing the periodical online pattern and instability issues of time encoding methods in streaming systems. Online A/B tests on Douyin Music App show +0.509% and +0.758% improvements in user active days and app duration respectively, while offline experiments also demonstrate the method's effectiveness. Interest Clock has been widely deployed in Douyin Music's online recommendation systems.

- Kim et al. (2024a) propose a domain- and data-agnostic item content representation learning framework for cold-start recommendations. The proposed method utilizes Transformer-based architectures to naturally fuse multimodal content signals and is end-to-end trainable without relying on human-labeled classification datasets. By training solely on user activities, the learned item representations better preserve fine-grained user preferences compared to models trained on proxy classification tasks. Extensive experiments on movie and news recommendation benchmarks demonstrate the framework's superior cold-start performance and generalizability across multiple domains, outperforming state-of-the-art baselines.

3.3.1 Music & Podcasts

- Zhang et al. (2024a) MaTrRec is a novel sequential recommendation model that combines the strengths of Mamba and Transformer architectures to address limitations of existing approaches. While Transformers struggle with long sequences due to quadratic complexity, and Mamba underperforms on short sequences, MaTrRec leverages Mamba's linear complexity for efficient long-range dependency modeling and Transformer's multi-head attention for short-range and global dependencies. The model architecture consists of an embedding layer, Mamba blocks, a Transformer encoder, feed-forward layers, and residual connections. Experiments on five public datasets (ML-1M, Musical, Health, Electronics, Office) demonstrate MaTrRec's superior performance over state-of-the-art baselines, with significant improvements in both long and short interaction sequences. Notably, it achieves up to 33% improvement on the highly sparse Amazon Musical Instruments dataset, effectively addressing the cold start problem. Ablation studies validate the contribution of each component, while analysis of dropout rates and maximum sequence length provides insights into optimal model configuration.

Figure 5: Concept map illustrating the advancements in Entertainment and Media recommender systems. The map is divided into two main areas: **Innovative Frameworks and Methodologies** and **Music & Podcasts**. Under each category, key research studies and methods are outlined, showing the progression and relationships between different approaches. Cross-connections indicate how certain methodologies influence or relate to others across categories, highlighting the interdisciplinary nature of recent advancements in recommender systems.

- Ferrante et al. (2024) demonstrates the feasibility of decoding music from cross-subject fMRI data using advanced computational approaches. Leveraging the GTZan fMRI dataset of 540 musical stimuli across 10 genres, the researchers employed the CLAP model to extract latent representations of musical stimuli and developed voxel-wise encoding models to identify music-responsive brain regions. They implemented three functional alignment techniques (anatomical, hyperalignment, linear) to address inter-subject variability, with the linear method achieving the highest identification accuracy of 0.9012 ± 0.01573. The study identified key music-processing areas in the brain, including the superior temporal gyrus, primary auditory cortex, planum temporale, and inferior parietal lobule. Genre classification showed high accuracy for classical and jazz, with some confusion between related genres like rock and metal. The decoding-in-time analysis revealed increasing identification accuracy over the stimulus duration, peaking towards the end of the 18-second window. This research not only advances our understanding of neural music processing but also has potential applications in personalized music therapy and recommendation systems. The study's limitations include the inherent noise in fMRI signals and challenges in decoding fine temporal aspects of music due to fMRI's low temporal resolution. Future work could explore higher temporal resolution neuroimaging methods and more sophisticated generative models to further advance the field of neuromusicology. CopyRetryClaude can make mistakes. Please double-check responses.

- Jing et al. (2024) introduces a Heterogeneity-aware Deep Bayesian Network (HDBN) for personalized music recommendation, addressing four key types of heterogeneity: emotion heterogeneity across users and within a user, and music mood preference heterogeneity across users and within a user. The model comprises four main components: (1) an inference network for personalized prior latent emotion distribution modeling, (2) another inference network for posterior latent emotion distribution modeling, (3) user grouping for capturing diverse music mood preferences, and (4) Bayesian neural networks for predicting music mood preferences. To validate the model, the authors constructed a large-scale dataset called EmoMusicLJ, containing 129,104 interactions from 12,557 users and 6,095 music tracks. Experimental results demonstrate that HDBN significantly outperforms state-of-the-art baseline models on Hit Rate (HR) and Normalized Discounted Cumulative Gain (NDCG) metrics. Ablation studies confirm the effectiveness of each component, with the music mood preference heterogeneity across users component showing the most significant impact. The model also demonstrates interpretability in learning meaningful latent emotion representations. This work advances emotion-aware music recommendation by explicitly modeling various heterogeneities, offering improved personalization and recommendation accuracy.

- Mei et al. (2024) investigates the impact of incorporating negative feedback in personalized music recommendation systems, using a hybrid SASRec-BERT4Rec transformer architecture. The authors demonstrate that utilizing explicit negative feedback (e.g., "thumbs-down") as hard negative samples during training significantly improves test accuracy by 6% compared to random negative sampling, while reducing training time by approximately 60%. They explore the relationship between the number of randomly sampled negatives and model performance, finding an optimal range before false negatives degrade accuracy. The study also shows that including implicit negative feedback (e.g., song skips) as input features increases user coverage and slightly improves accuracy. Analysis of feedback embeddings reveals that skip behavior is more similar to explicit negative feedback than positive feedback. The authors examine the trade-off between maximizing true positives and minimizing false positives, noting that different proportions of hard negatives may be optimal for candidate generation versus ranking tasks. Experiments were conducted on three datasets: Piki, Spotify's Sequential Skip Prediction, and a proprietary Pandora dataset, with consistent results across varying proportions of feedback types. This work provides valuable insights into leveraging negative feed-

back for enhancing music recommendation systems' performance and coverage.

- (Moradi et al., 2024) This paper addresses cultural and popularity biases in music recommendation systems by enhancing the Prototype-based Matrix Factorization (ProtoMF) method. The authors propose two novel techniques: Prototype K-filtering, which uses only the k nearest prototypes to generate user and item representations, and a Prototype-Distributing Regularizer, which encourages a more uniform distribution of prototypes in the embedding space. Using the LastFM-2b and MovieLens-1M datasets, they demonstrate that these techniques significantly reduce bias without compromising recommendation quality. The study reveals that in the baseline ProtoMF model, prototypes cluster around popular items, leading to unfair representation of less popular and culturally diverse content. The proposed enhancements mitigate this issue by improving the visibility of underrepresented items and cultures. Experimental results show that the combined approach outperforms existing methods in terms of fairness metrics while maintaining competitive performance in standard recommendation metrics like HitRatio@10 and NDCG@10. The authors emphasize the importance of addressing cultural bias in recommendation systems to promote diversity and inclusivity in music consumption. They also discuss ethical considerations and potential limitations of their approach, highlighting the need for ongoing research in this area.

- Dai et al. (2024) introduce ReCODE, a novel model-agnostic framework for modeling repeat consumption in recommender systems using neural ordinary differential equations (ODE). The authors address the limitations of existing methods that rely on heuristic assumptions about temporal dynamics of repeat consumption by leveraging the flexibility of neural ODEs to capture complex patterns without distribution assumptions. ReCODE comprises two key components: a static recommendation module for capturing basic user preferences, and a dynamic repeat-aware module that uses neural ODEs to model temporal patterns of repeat consumption. The framework can be integrated with various existing recommendation models, including collaborative-based and sequential-based approaches. Experiments on two real-world datasets (MMTD and Nowplaying-RS) demonstrate that ReCODE significantly outperforms state-of-the-art baselines, including GRU4Rec, Caser, NARM, SASRec, and ContraRec, across multiple evaluation metrics (Recall@K and NDCG@K). The authors show that ReCODE consistently improves performance when applied to different base models (MF, NCF, GRU4Rec, SASRec) and exhibits increased effectiveness as repeat ratios in datasets increase. This work highlights the importance of accurately modeling repeat consumption in real-world recommender systems, particularly in domains with high repetition rates such as music streaming.

- Zhu et al. (2024a) presents a comprehensive tutorial on multimodal pretraining and generation techniques for recommendation systems. The authors outline three key areas: multimodal pretraining, multimodal generation, and industrial applications with open challenges. The tutorial covers self-supervised pretraining paradigms, multimodal pretrained models (e.g., CLIP, GPT-4), and their adaptation to recommendation tasks. It discusses various pretraining approaches for sequence, text, audio, and multimodal data in recommendation contexts. The generation section explores text and image generation techniques, including personalized approaches. The tutorial also highlights successful industrial applications and open challenges, such as multimodal representation fusion and efficient adaptation of large language models. The authors emphasize the importance of leveraging multimodal data to enhance recommendation performance, particularly in multimedia services. They provide insights into recent advancements like using neural ordinary differential equations for modeling repeat consumption patterns and implementing personalized content generation. The tutorial aims to bridge the gap between multimodal learning and recommender systems research, offering valuable insights for both academic researchers and industry practitioners in this rapidly evolving field.

- Marey et al. (2024) introduces PACE (PAttern-based user Consumption Embedding), a novel framework for modeling activity-driven music listening behaviors by leveraging regular consumption patterns. PACE encodes user consumption histories as multivariate time series across four channels: volume, repetition, organicity, and liked content. It employs dictionary learning to detect stereotypical listening behaviors, represented as atoms, and projects user histories onto these atoms to generate interpretable user embeddings. The framework is evaluated through an activity prediction task using survey data from Deezer users, focusing on six common activities: waking up, commuting, working, sports, socializing, and falling asleep. PACE embeddings achieve intermediate performance between simple baselines and a strong activity-based baseline, demonstrating their ability to capture meaningful listening patterns. The method shows particular strength in predicting more regular activities like waking up and falling asleep. Analysis of the learned atoms reveals interpretable weekly listening patterns consistent with expected behaviors for different activities. The authors suggest that PACE's ability to capture regular listening behaviors makes it promising for improving contextual music recommendation and understanding soundtracking practices. Future work could involve integrating content-based information and exploring applications in recommender systems.

- Nir et al. (2024) presents VCR (Video representation for Contextual Retrieval), a novel approach for efficient content discovery in large video archives. The method leverages multimodal video insights, including automatic speech recognition (ASR), optical character recognition (OCR), and frame captioning, to create a unified text-based video representation. This representation is then embedded into a semantic space using either fine-tuned language models (BERT, RoBERTa, DeBERTa) or OpenAI's GPT embeddings. The system introduces a Topics-Map visualization tool that allows users to explore video content through an interactive 2D semantic map of topics. VCR achieves state-of-the-art results on unsupervised Text-Video Retrieval tasks over the MSR-VTT dataset without fine-tuning. The authors evaluate their method on a newly introduced TED Topics Dataset and conduct a user study demonstrating significant improvements in content discovery compared to baseline methods. The system's effectiveness is quantified using Mean Reciprocal Rank (MRR) and Recall@k metrics, with the multimodal approach consistently outperforming ASR-only representations. The paper contributes to the field of video archive exploration by offering a scalable, domain-adaptable solution that combines advanced AI models with an intuitive user interface, enabling more efficient and effective content discovery in large media archives.

- Joung and Lee (2024) presents a novel approach to improve music auto-tagging performance in noisy environments by leveraging Domain Adversarial Training (DAT) to create robust music representations. The proposed method integrates a three-step training process: pretraining the Feature Extractor (FE), pretraining the Domain Classifier (DC), and finetuning the FE while training the Label Predictor (LP). The architecture employs CLMR with SampleCNN as the encoder for the FE. The study utilizes diverse datasets including MTG-Jamendo, MagnaTagATune (MTAT), Audioset, and Musan to simulate real-world noise conditions. The model's performance is evaluated using AUC and AP metrics across various Signal-to-Noise Ratio (SNR) settings. Results demonstrate that the proposed method, especially when incorporating additional unlabeled noisy data, enhances generalization and robustness in music auto-tagging tasks. The model shows consistent performance across different noise conditions, including tests on the Musan noise dataset, affirming its resilience to various noise types. This approach offers a promising solution for improving music auto-tagging in real-world, noisy scenarios, particularly beneficial for video-streaming platforms where music is often mixed with environmental sounds.

- Beattie et al. (2023) present a novel evaluation framework for assessing attribute association bias in latent factor recommendation algorithms. The authors introduce methods to detect and quantify how sensitive attributes like

gender can become entangled in trained recommendation latent spaces, potentially leading to stereotyping and representation harms. The framework includes four key components: latent space visualization, bias direction computation, bias evaluation metrics, and classification for explaining bias. The authors demonstrate the framework's effectiveness through a case study on user gender bias in podcast recommendations, using a production-level deep neural network model. The study reveals significant levels of gender bias persisting even after removing gender as an explicit feature, suggesting the presence of implicit, systematic bias. The framework successfully identifies and measures attribute association bias, showing its potential for uncovering stereotyped relationships in recommendation systems. The authors also discuss the challenges of mitigating systematic bias and the need for careful consideration of when and how to address such biases. This work contributes to the growing field of fairness in recommender systems by providing a comprehensive approach to evaluating representation bias, which has been largely unexplored compared to other forms of bias like popularity or exposure bias.

- McDonald et al. (2023) introduces the "impatient bandit" algorithm for optimizing long-term user engagement in recommender systems with delayed rewards. The approach combines a Bayesian reward model that leverages progressively revealed intermediate outcomes with Thompson sampling to balance exploration and exploitation. The reward model uses historical data to learn prior and noise covariance structures through a meta-learning approach, enabling rapid inference about new items. The algorithm is applied to a podcast recommendation problem, aiming to maximize user engagement over a 60-day period. Experimental results on real-world data from Spotify demonstrate that the proposed method significantly outperforms baselines that rely on short-term proxies or wait for full reward realization. The impatient bandit shows superior performance in identifying high-quality content quickly, adapting to changing content libraries, and maintaining broader exploration of the action space. The authors also discuss potential extensions to personalized recommendations and other domains like hyperparameter optimization. The paper contributes a novel approach to addressing the challenge of optimizing for long-term goals in recommendation systems while efficiently handling the trade-off between rapid learning and alignment with long-term objectives.

- Chen et al. (2023b) introduces FMMRec, a novel fairness-aware multimodal recommendation approach that addresses the challenge of sensitive information leakage in multimodal content. The method employs a three-phase process: pretraining, disentanglement learning, and modality-guided fairness learning. In the disentanglement phase, it generates biased and filtered modal embeddings by maximizing and minimizing sensitive attribute prediction ability, respectively. The modality-guided fairness learning phase utilizes these disentangled embeddings to mine fair and unfair user-user structures, which are then used to enhance user representations. FMMRec employs adversarial learning with role indicator embeddings to eliminate sensitive information from both user and item representations. Experiments on MovieLens and MicroLens datasets demonstrate FMMRec's superior performance in achieving counterfactual fairness while maintaining competitive accuracy compared to state-of-the-art baselines. The method effectively reduces sensitive information leakage in both explicit and implicit user representations across multiple attributes (gender, age, occupation). The ablation study confirms the effectiveness of each component, particularly the importance of both fair and unfair structure mining and the role indicator embeddings. FMMRec represents a significant advancement in addressing fairness issues in multimodal recommendation systems.

3.4 Health Care

The healthcare industry is experiencing a significant transformation through the integration of advanced artificial intelligence and machine learning techniques in recommender systems. These innovations are addressing critical challenges across various aspects of healthcare delivery and management. Novel approaches like the Empirical Soft Regret (ESR) loss function and models such as LEADER and DKINet are enhancing the accuracy

and efficiency of treatment and medication recommendations. Concurrently, there's a growing focus on fairness and privacy in healthcare AI, exemplified by frameworks like RAREMed and F2PGNN, which aim to provide equitable recommendations for patients with rare diseases while preserving data privacy. AI-assisted tools like FAIIR are revolutionizing mental health support services, while sophisticated emotion recommender systems are improving patient experience analysis. The field is also witnessing advancements in resource allocation and optimization, with innovative approaches to cancer treatment center recommendations and hospital queue management. Furthermore, the exploration of human-ML collaborative systems is paving the way for more effective integration of AI in clinical decision-making processes. These developments collectively promise to enhance patient care, optimize healthcare resources, and support more informed clinical decisions while addressing crucial ethical considerations in AI-driven healthcare.

- Tan and Frazier (2024) propose the Empirical Soft Regret (ESR) loss function for training machine learning models in the black-box predict-then-optimize setting, where only the reward for the chosen action is observed. ESR targets minimizing the regret from taking suboptimal actions based on the model's predictions. The authors prove that under certain conditions, using ESR achieves asymptotically optimal regret. Experiments on semi-synthetic and real-world datasets demonstrate that ESR outperforms state-of-the-art methods in contextual bandits and conditional average treatment effect estimation.

- Obadinma et al. (2024) present the development and evaluation of FAIIR (Frontline Assistant: Issue Identification and Recommendation), an ensemble of domain-adapted and fine-tuned transformer models that leverages natural language processing to identify issues youth may be experiencing during text-based conversations with crisis responders. FAIIR achieved 94% accuracy and 64% F1-score on a dataset of over 700,000 unique conversations. The tool demonstrated unbiased performance across demographic subgroups and strong generalization in silent testing. Expert assessment validated FAIIR's predictions, showing higher agreement with the tool than original issue labels. The study highlights the potential of AI-assisted issue identification in enhancing youth mental health support services.

- Sühr et al. (2024) present a novel dynamic framework for the deployment of machine learning (ML) models in a performative, human-ML collaborative system. The authors introduce the collaborative characteristic function, which maps the performance of ML models to the performance of the human+ML system, and define collaborative learning paths characterizing possible dynamic deployment strategies. They conduct an empirical user study with 1,408 participants solving knapsack problems, showcasing how the dynamic process can converge to different stable points. The results suggest that for many levels of ML performance, the equilibrium performance of the human+ML system is around 92% of the optimal solution. The authors also find that monetary incentives do not improve human decision quality, and humans sometimes make decisions worse than ML recommendations. The paper highlights the importance of adopting a dynamic approach when deploying ML models that interact with human decisions.

- Zhao et al. (2024) propose RAREMed, a novel medication recommendation model designed to address the fairness issue in recommender systems, where patients with rare diseases often receive less accurate recommendations compared to those with common conditions. RAREMed employs a transformer encoder with a unified input sequence approach to capture complex relationships among disease and procedure codes, and introduces two self-supervised pre-training tasks to learn specialized medication needs and interrelations among clinical codes. Experimental results on two real-world datasets demonstrate that RAREMed provides accurate drug recommendations for both rare and common disease patients, effectively mitigating unfairness in medication recommendation systems.

- Liu et al. (2024a) propose LEADER, a novel approach that leverages large language models (LLMs) for medication recommendation. LEADER addresses the out-of-corpus problem and high inference costs associated with

LLMs by modifying the output layer, fine-tuning the loss function, and introducing a feature-level knowledge distillation technique to transfer the LLM's capabilities to a smaller, more efficient model. The proposed student model also incorporates profile alignment to handle single-visit patients effectively. Extensive experiments on two real-world datasets demonstrate that LEADER outperforms state-of-the-art medication recommendation models in terms of effectiveness and efficiency.

- Murray et al. (2024) present a comprehensive analysis of patient narratives from the Australian healthcare review website Care Opinion, employing topic modeling and sentiment analysis techniques. The study reveals three main themes in patient experiences: clinical care, patient experience, and healthcare logistics. Notably, the authors find that extremely positive and negative sentiments are more strongly associated with subjective patient experiences than clinical outcomes. The analysis also uncovers important relationships between patient emotions and specific aspects of care. For instance, feelings of neglect and dismissal are closely linked to severe negative emotions like suicidality and hopelessness, while gratitude is associated with feeling cared for and involved. Based on these findings, the authors develop a novel probabilistic emotion recommender system using topic modeling in a Naive Bayes approach, which outperforms baseline models in predicting emotions and sentiments from patient narratives. The proposed methodology offers an interpretable and cost-effective way to harness unconstrained patient feedback, complementing traditional patient experience surveys. To facilitate the adoption of this approach in healthcare research and practice, the authors provide an R package and an online dashboard.

- Liu et al. (2024b) propose DKINet, a novel framework that incorporates domain knowledge from the Unified Medical Language System (UMLS) with clinical manifestations from electronic health records (EHR) to improve medication recommendation. The key contributions include: 1) a graph aggregation module to extract informative knowledge from the UMLS graph; 2) a knowledge-injected patient representation module that integrates domain knowledge based on clinical manifestations; and 3) a historical medication-aware patient representation module to capture the longitudinal influence of historical medications. Extensive experiments on three benchmark datasets (MIMIC-III, MIMIC-IV, and eICU) demonstrate the superiority of DKINet over state-of-the-art methods. The proposed approach provides a new perspective on leveraging external domain knowledge to enhance medication recommendation, potentially assisting physicians in making more accurate prescriptions for patients with complex health conditions.

- Unnikrishnan et al. (2024) propose TreatmentRecommender, an algorithm for predicting the expected improvement in a patient's condition for each of several alternative treatments, where the treatment data come from a multi-armed randomized clinical trial (RCT). The key challenges addressed are: 1) missing rationale for treatment assignment in the RCT data; 2) missing verification evidence when predicting outcomes for treatments not assigned to patients; and 3) missing evidence due to uneven distribution of patients among treatments. The authors introduce counterfactual treatment verification to deal with missing verification evidence and use an ensemble of therapy-level models to boost evidence for outcome prediction per treatment. They devise a validation procedure based on alignment between the treatment assignments of the RCT and the top-ranked treatment recommended by the algorithm for each patient. Results demonstrate that the approach effectively leverages RCT data for learning and validation, showing that the recommended treatments improve outcomes. The study provides a basis for establishing decision support routines on treatments tested in RCTs but not yet deployed clinically.

- Seidi (2024) proposes a novel system for recommending the most suitable cancer treatment center to high-risk patients based on their income and location. The system consists of two main components: 1) a survival analysis model to calculate the risk score for each patient and select those above a prede-

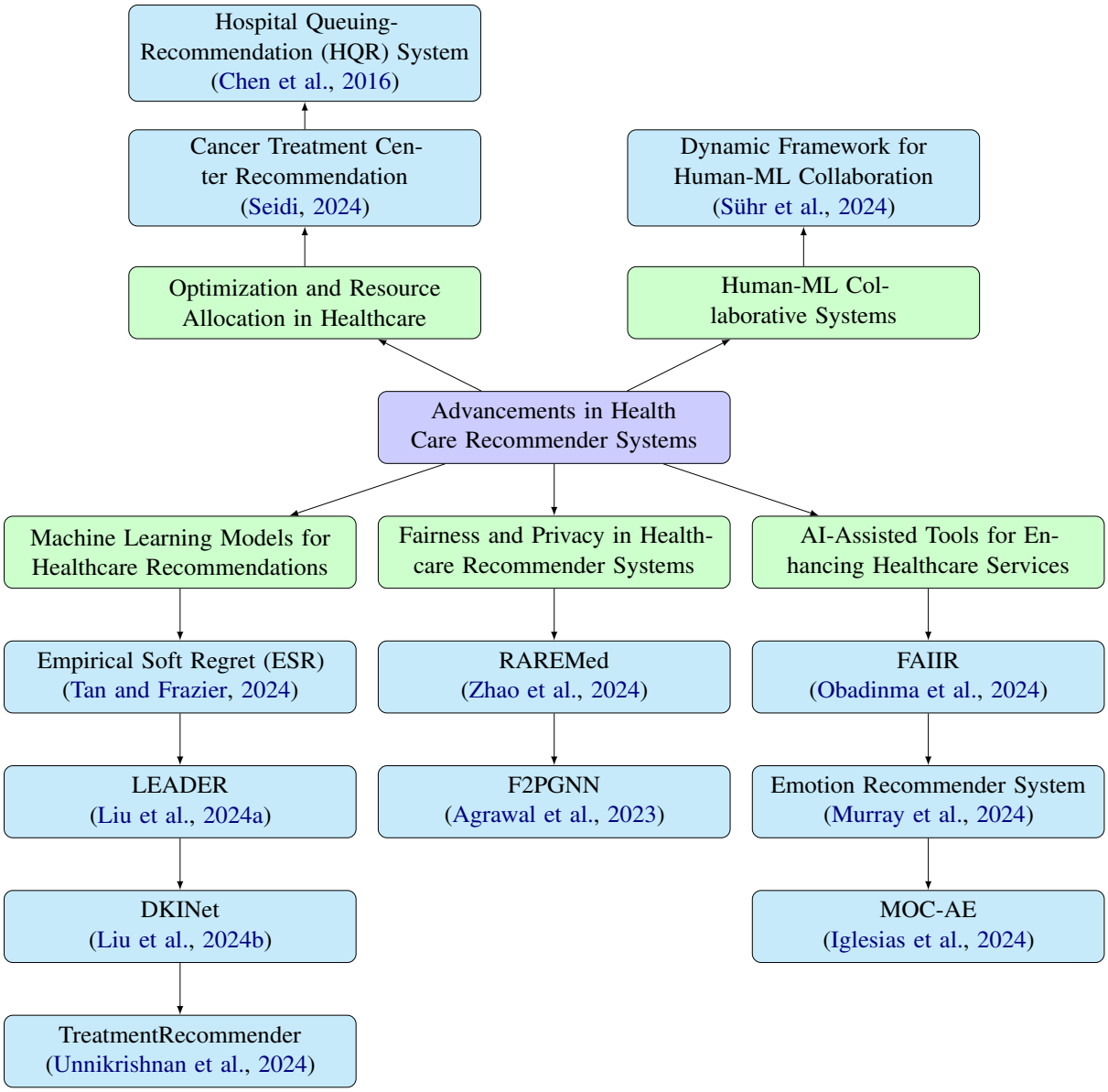

Figure 6: Concept map illustrating the advancements in Health Care recommender systems. The map is divided into five primary categories: **Machine Learning Models for Healthcare Recommendations**, **Fairness and Privacy in Healthcare Recommender Systems**, **AI-Assisted Tools for Enhancing Healthcare Services**, **Optimization and Resource Allocation in Healthcare**, and **Human-ML Collaborative Systems**. Under each category, key research studies and methods are outlined, showing the progression and relationships between different approaches.

fined threshold; and 2) a stable matching algorithm to assign eligible patients to the best cancer treatment center considering their location and income. The problem is formulated as a graph-theoretic one, with patients and cancer centers represented as vertices and edges connecting them based on affordability and accessibility criteria. The algorithm iteratively matches patients to centers until all available staffed beds are filled. Experimental results using patient data from the SEER dataset and NCI-designated cancer center information demonstrate the effectiveness of the approach in assigning high-risk patients to the most appropriate treatment facilities. This work highlights the potential of leveraging advanced patient management systems and smart city technologies to optimize cancer care delivery and resource allocation.

- Iglesias et al. (2024) propose a novel deep learning architecture, Multi-Output Classification Autoencoder (MOC-AE), for content-based image retrieval (CBIR) in medical diagnosis. The model extracts both anatomical and pathological features from brain MRI scans to retrieve the most similar cases from a database, aiding clinicians in their decision-making process. The key contributions include: 1) the ability to generate enriched image descriptors using only binary tumor presence labels, eliminating the need for costly tumor segmentation; 2) a dual-objective learning scheme that balances the representation of healthy and abnormal features; and 3) state-of-the-art performance in retrieving similar cases, outperforming previous approaches with a Dice coefficient of 0.474 for both tumoral and healthy regions. The authors demonstrate the model's potential for improving the efficiency and accuracy of comparative diagnostics and treatment of tumoral pathologies while reducing costs. The flexible architecture can be extended to various medical diagnostic cases, marking a significant advancement in AI-assisted image retrieval for clinical decision support systems.

- Seidi (2024) propose FD-GATDR, a novel federated-decentralized learning framework for doctor recommendation using electronic health records (EHRs). The key contributions include: 1) constructing a heterogeneous graph to capture structured information from EHRs; 2) developing a heterogeneous graph attention network (HGAT) that considers time sensitivity and node heterogeneity for patient, doctor, and service representation; and 3) introducing a federated decentralized learning algorithm to address data sharing privacy concerns among hospitals. The proposed model outperforms baseline methods by up to 6.2% in AUC on an EHR dataset, while the federated-decentralized learning algorithm achieves performance comparable to a fictitious fusion center with a convergence rate of O(1/T). This work marks a significant step towards leveraging graph learning and federated learning techniques for personalized doctor recommendation while preserving patient privacy.

- Agrawal et al. (2023) propose F2PGNN, a novel framework for achieving fairness in federated graph-based recommender systems while preserving user privacy. The key contributions include: 1) an inductive graph expansion algorithm that incorporates higher-order interactions while minimizing communication overhead; 2) a fairness-aware optimization strategy that mitigates group unfairness without compromising user privacy; and 3) enhanced privacy protection through local differential privacy techniques applied to model updates and group statistics. Experimental results on three datasets demonstrate that F2PGNN effectively reduces group unfairness by 47-99% compared to the state-of-the-art method while maintaining recommendation performance and privacy. The proposed framework offers a promising solution for achieving equitable and personalized recommendations in federated learning environments.

- Chen et al. (2016) propose a Patient Treatment Time Prediction (PTTP) algorithm and a Hospital Queuing-Recommendation (HQR) system to minimize patient wait times and improve hospital queue management. The key contributions include: 1) developing the PTTP algorithm based on an improved Random Forest to predict the waiting time for each treatment task; 2) introducing the HQR system to

recommend efficient and convenient treatment plans for each patient based on predicted waiting times; and 3) parallelizing the PTTP algorithm and HQR system on Apache Spark for efficiency and scalability. Extensive experiments using real hospital data demonstrate the superiority of the proposed approach, with the PTTP algorithm achieving high accuracy and robustness, and the HQR system effectively reducing average patient waiting times. This work presents a promising solution for intelligent patient queue management and treatment recommendation in a big data healthcare environment.

3.5 Finance

Recommender systems are revolutionizing financial trading and investment decision-making. These systems now provide portfolio managers (PMs) with ranked buy and sell recommendations, complete with conviction levels and track records, integrating seamlessly with risk management and portfolio software. Explainable AI advancements are driving RS adoption in finance, allowing PMs to maintain control while utilizing AI insights. Key trends include enhanced data visualization and tools targeting discretionary managers.

In the NFT market, a graph-based RS leverages transaction, image, text, and price data to overcome limited feedback and user anonymity. An interpretable decision-making model in the Pakistan Stock Exchange uses technical indicators and SHAP to boost investor portfolios. PicPay's hybrid RS, combining probabilistic and gradient boosting filters, shows notable conversion rate improvements.

ZeroShotALI, utilizing SentenceBERT and GPT-4, excels in zero-shot auditing of financial reports. Adaptive Collaborative Filtering (ACF) captures time-dependent user preferences, outperforming benchmarks. Knowledge graph-driven RS in finance use reinforcement learning and XGBoost for explainable recommendations. NFT-MARS addresses sparse interactions and the dual nature of NFTs, outperforming various baselines.

Mean-Variance Efficient Collaborative Filtering (MVECF) balances risk and return in portfolio recommendations. A punishment-based algorithm incentivizes high-quality content in online RS. LinkedIn's Account Prioritizer, an AI-driven tool for sales account prioritization, significantly enhances renewal bookings and demonstrates the impact of AI on sales strategies.

- Vidler (2024) explores the application of Recommender Systems, a form of AI, in financial trading and investment decision-making. The key points include: RS can replicate the roles of human analysts in providing ranked lists of high-conviction buy and sell recommendations to portfolio managers (PMs). RS outputs include metadata such as conviction levels, track records, and explanations to help PMs interpret and trust the recommendations. PMs use the RS recommendations in conjunction with risk management systems, portfolio management software, and alpha analysis programs for pre-trade analysis and decision support. The adoption of RS in finance is increasing, driven by advancements in explainable AI and the need for PMs to leverage AI while maintaining control over investment decisions. Trends in RS development include improved data visualization, targeting discretionary managers, and bridging the gap between non-technical investment staff and powerful AI tools.

- Choi et al. (2024) propose a recommender system for NFT collectibles that leverages a graph-based approach and incorporates item features to address the challenges of limited feedback information and user anonymity in the NFT market. The system utilizes transaction data, image features, textual information, and price data to generate precise recommendations tailored to individual preferences. The authors develop a data-efficient graph-based recommender system that captures the complex relationships between items and users, generating node embeddings that incorporate both node feature information and graph structure. Experiments on real-world NFT transaction data demonstrate that the proposed graph-based recommender system significantly improves performance after incorporating all types of item features as side information, outperforming baseline models. This study highlights the importance of leveraging item features and graph-based models to overcome the limitations of NFT data and provide accurate recommendations in the growing NFT market.

- Arshad et al. (2023) present an interpretable

decision-making model for investment recommendations in the Pakistan Stock Exchange (PSX). The proposed solution incorporates technical indicators and the SHAP explainability technique to provide valuable insights into the factors influencing forecasted recommendations for investors with varying trading strategies. A case study demonstrates the model's effectiveness in enhancing an investor's portfolio value. The results highlight the importance of integrating interpretability in financial forecasting models to boost stakeholder confidence and transparency in the stock exchange domain.

- Mendonça et al. (2023) present a Switching Hybrid Recommender System employed by the fintech PicPay to effectively promote strategic financial services to its 30 million monthly active users. The system combines a Probabilistic Filter for new products with limited historical data and a Gradient Boosting Filter that uses machine learning to predict user interest for items with more conversion data. A/B test results demonstrate uplifts of up to 3.2% in conversion rates compared to the default recommendation strategy, highlighting the effectiveness of the hybrid approach in promoting items while maintaining a positive user experience.

- Hillebrand et al. (2023) introduce ZeroShotALI, a novel recommender system that combines a domain-specifically optimized SentenceBERT model with the GPT-4 large language model to match relevant text segments from financial reports to legal requirements in accounting standards, enabling zero-shot auditing of new reports. Evaluating against several strong baselines, they find that the two-stage approach of SentenceBERT retrieval followed by GPT-4 filtering outperforms other methods, highlighting the benefit of leveraging domain-specific solutions augmented with state-of-the-art LLMs for this task. The authors note the system could be further improved through domain-specific LLM fine-tuning, advanced prompting techniques, and expansion to assess requirement completeness.

- Ghiye et al. (2023) propose Adaptive Collaborative Filtering (ACF), a time-dependent recommender system for financial products that adaptively discounts past client-product interactions using personalized decay functions. They argue that interest in financial products fades over time at different rates for different clients. To capture this, ACF learns personalized weighting coefficients to model the varying utility of past interactions based on the user-item pair and time elapsed. Evaluating on a proprietary dataset from BNP Paribas, ACF with learned exponential decay rates outperforms state-of-the-art benchmarks, demonstrating the importance of explicitly modeling the dynamic nature of user preferences in financial recommendation settings. The authors note future work could explore higher-order graph connectivity and dynamic node features.

- Verma et al. (2023) propose two knowledge graph (KG)-driven recommender systems for personalized article recommendation in financial services, using reinforcement learning and XGBoost. They automatically generate KGs from both structured and unstructured data, and employ KG embeddings from TuckER and TransE. The RL-based approach uses an MDP framework to learn an exploratory policy over the KG, generating explainable recommendations via path reasoning. The XGBoost approach incorporates KG embeddings as features. Evaluating on proprietary data, the RL method outperforms XGBoost and baselines, providing accurate recommendations and interpretable reasoning paths. Post-hoc explanations for XGBoost are generated using SHAP and ELI5. Results highlight the potential of combining advanced ML with KGs for explainable recommendation in finance.

- Kim et al. (2023) propose NFT-MARS, the first recommender system designed to address the unique challenges of the rapidly growing NFT market, such as extremely sparse user-item interactions, anonymity, and the dual nature of NFTs as both artwork and financial assets. NFT-MARS employs (1) graph attention to handle sparse interactions, (2) multimodal attention to incorporate user-specific feature preferences, and (3) multi-task learning to consider both artistic and investment

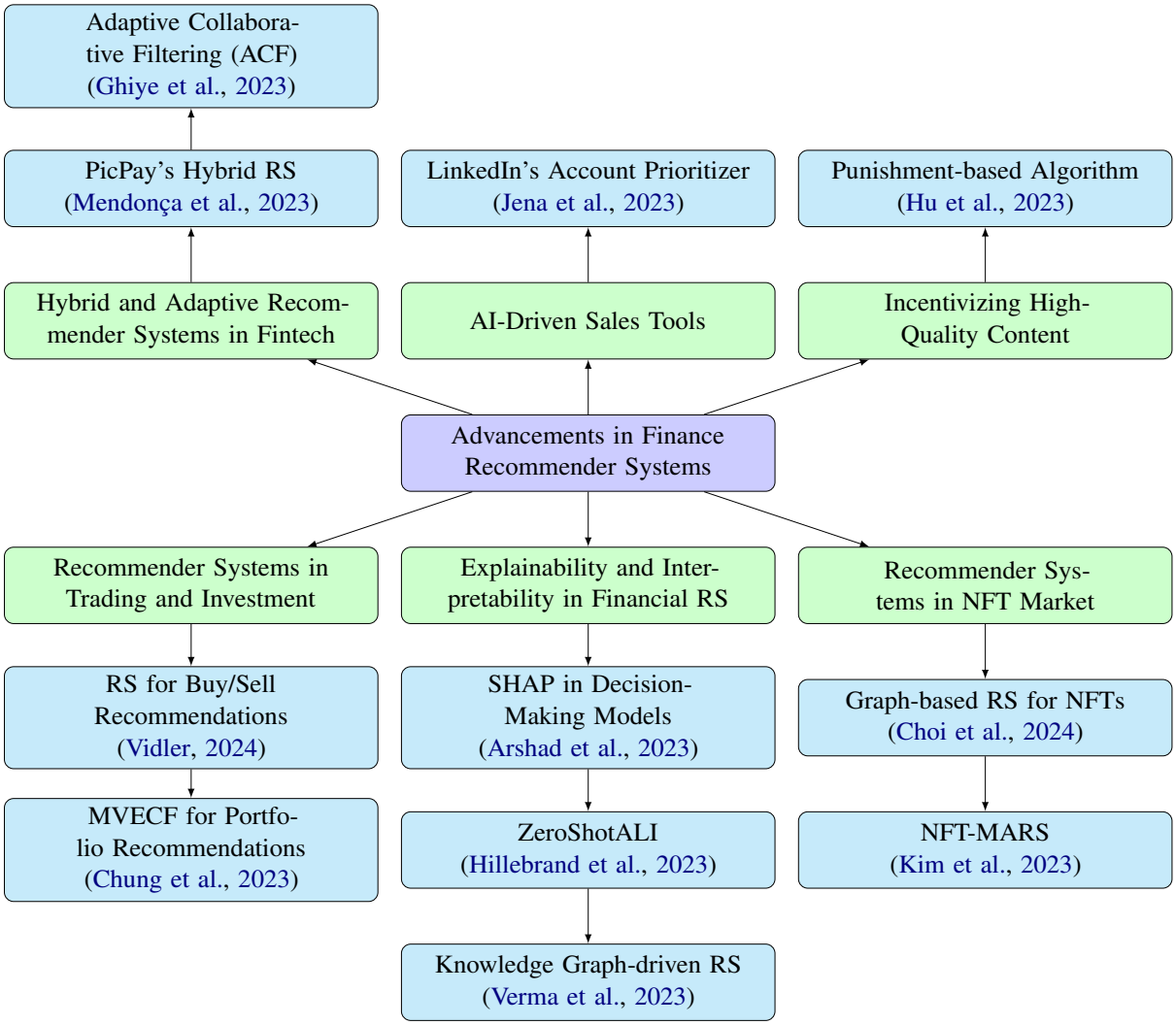

Figure 7: Concept map illustrating the advancements in Finance recommender systems. The map is divided into six primary categories: **Hybrid and Adaptive Recommender Systems in Fintech**, **AI-Driven Sales Tools**, **Incentivizing High-Quality Content**, **Recommender Systems in Trading and Investment**, **Explainability and Interpretability in Financial RS**, and **Recommender Systems in NFT Market**. Under each category, key research studies and methods are outlined, showing the progression and relationships between different approaches. Cross-connections indicate how certain methodologies influence or relate to others across categories, highlighting the interdisciplinary nature of recent advancements in finance recommender systems.

aspects. Evaluating on real-world transaction data from four popular NFT collections, NFT-MARS significantly outperforms various baselines. Ablation studies and attention score analysis demonstrate the effectiveness of the key components and reveal varying feature importance across collections. The authors highlight the potential of tailored recommender systems in driving the future growth of the NFT market.

- Chung et al. (2023) propose Mean-Variance Efficient Collaborative Filtering (MVECF) for personalized stock recommendation, addressing the unique challenges of incorporating both user preferences and risk-return characteristics. MVECF modifies the weighted matrix factorization (WMF) model using regularization to improve the Pareto optimality between risk (return variance) and return (mean return). The regularization term is then restructured into an ordinary WMF form for computational efficiency. Experiments on real-world mutual fund and stock ownership data show that MVECF significantly improves the mean-variance efficiency (measured by Sharpe ratio) of recommended portfolios while sacrificing minimal recommendation accuracy compared to state-of-the-art models. Furthermore, MVECF can be easily integrated into graph-based ranking models through a modified sampling scheme. This work highlights the importance of tailoring recommender systems to the specific needs of the financial domain.

- Hu et al. (2023) study the impact of a platform's learning algorithm on the quality of user-generated content in online recommender systems. They show that classical online learning algorithms like LinHedge and LinEXP3 incentivize producers to create low-quality content, with equilibrium quality decreasing over time for typical learning rate schedules. Motivated by these negative results, the authors propose an alternative punishment-based algorithm that incentivizes high-quality content creation by excluding producers below a quality threshold. For one dimension, this approach achieves near-optimal content quality (compared to a heuristic benchmark) across almost all time steps. Increasing the learning rate alone leads to non-existence of symmetric pure strategy equilibria. The work highlights the unintended consequences of learning algorithms on content quality and opens the door to designing algorithms that promote high-quality content.

- Jena et al. (2023) present Account Prioritizer, a data product developed at LinkedIn to automate sales account prioritization. It uses machine learning recommendation models to predict upsell opportunities and churn risks at the account and product level. The models are trained on historical account and product bookings data and retrained monthly. Integrated with CrystalCandle, an explainable AI system, Account Prioritizer provides sales representatives with instance-level explanations to guide outreach efforts and build trust. An A/B test showed an 8.08% increase in renewal bookings from using Account Prioritizer. The authors also conducted a causal study using coarsened exact matching, finding a 20.4% lift for sales reps who consistently used the system. Account Prioritizer demonstrates the potential for AI-driven sales prioritization and the importance of explainability in driving adoption and impact.

3.6 Travel & Tourism

The travel and tourism industry is experiencing a revolutionary transformation driven by advanced artificial intelligence and machine learning technologies in recommender systems. Recent innovations are addressing complex challenges in personalized travel experiences, crowd management, and dynamic transportation optimization.

Cutting-edge frameworks like SBTREC and BTREC are leveraging transformer-based architectures to analyze user behavior and preferences, enhancing the accuracy of personalized tour recommendations. These models are pushing the boundaries of natural language processing in tourism, with systems like zIA demonstrating the potential of AI-powered chatbots to provide engaging, context-aware travel assistance.

Concurrently, there's a growing focus on developing configurable and context-aware Travel Recommender Systems (TRSs) to adapt to diverse operational environments. Algorithms like SCAIR are tackling the complex issue of crowd management in popular tourist destinations, optimizing

itineraries for both individual satisfaction and overall system efficiency.

In transportation, novel approaches such as the passenger route and departure time guidance system for urban rail transit are enhancing resilience during disruptions. Advanced techniques like non-stationary multi-armed bandits are being applied to predict tourism demand, while personalized route planning algorithms are integrating individual preferences with global traffic optimization.

These developments are not only improving the quality of travel recommendations but are also reshaping how tourists interact with destinations, fostering more efficient, enjoyable, and sustainable travel experiences. As the field progresses, it promises to deliver increasingly sophisticated, adaptive, and user-centric solutions, while also addressing crucial challenges in scalability, real-time data management, and ethical considerations in AI-driven tourism.

- Pereira et al. (2024) investigates the configurability of Travel Recommender Systems (TRSs) across different operational contexts. The authors analyzed 40 primary studies, focusing on algorithms, data types, outcomes, and configuration support. They found that hybrid-based approaches are most common, combining traditional filtering techniques to handle complex scenarios. While real-time data usage is increasing, most TRSs still rely on historical data, limiting their ability to adapt to dynamic contexts. The study reveals a significant lack of configuration support for TRS providers, with existing systems primarily focused on customizing user experiences rather than adapting to diverse operational environments. The authors identify key challenges, including the need for tailored solutions to handle specific contexts (e.g., urban vs. natural park settings), difficulties in managing real-time data, and the absence of a common methodology for configuring critical TRS components. They propose potential future directions, such as leveraging software product line engineering, low-code development platforms, and AI-based tools to facilitate the design-to-deployment process for TRS providers. The study concludes that current TRSs are largely inflexible, off-the-shelf solutions, highlighting the need for more configurable and context-aware systems to meet the diverse requirements of different travel scenarios.

- Zhuo et al. (2024) addresses the challenge of guiding passengers during disruptions in urban rail transit (URT) networks. The authors propose a three-feature four-group passenger classification principle, categorizing passengers based on temporal, spatial, and spatio-temporal characteristics. A mixed integer programming model is developed to optimize passenger routes and departure times at the network level, incorporating a First-in-First-out (FIFO) rule for oversaturated conditions. To handle large-scale problems, a two-stage solution approach is introduced, comprising Passenger Classification and Passenger Updating stages. The model is validated using both a small-scale artificial network and the Beijing Subway network. Results demonstrate significant reductions in total passenger travel time: 29.7% for the small case and 50% for the large-scale case, compared to scenarios without travel rescheduling guidance. The study also conducts sensitivity analyses on parameters such as train capacity, departure time window length, and disruption duration. The proposed approach shows superior computational efficiency compared to standard solvers, solving large-scale problems in minutes where traditional methods fail. This research contributes to enhancing the resilience and operational reliability of URT systems by providing timely, targeted guidance to passengers during disruptions.

- Selvaraj et al. (2024) presents a novel approach to personalized and context-aware route planning for vehicles using a combination of Graph Neural Networks (GNNs) and Deep Reinforcement Learning (DRL). The framework analyzes drivers' historical trajectories to classify driving behavior and associate it with road attributes, identifying individual preferences. It represents the road network as a graph structure, with GNNs modeling topology and attributes, while DRL optimizes route selection based on driver satisfaction, travel time, and global traffic flow. The system employs a two-step learning process: first training a generic model for traffic optimization, then fine-tuning it to incorporate

driver preferences. Evaluation on a real-world road network from the Luxembourg SUMO Traffic scenario demonstrates up to 17% improvement in selecting preferred routes compared to a generic model, and up to 46% reduction in travel time relative to a shortest distance-based approach. The framework shows flexibility in accommodating multiple driver preference variations and efficiently computes personalized routes in near real-time (average 0.16 seconds per request). This approach offers a promising solution for integrating individual preferences into intelligent transportation systems, potentially enhancing both traffic efficiency and user satisfaction in urban environments.

- Cassani et al. (2024) presents zIA, a generative AI-powered chatbot designed to assist tourists in Italy, specifically in the Molise region. The system leverages large language models and embedding techniques to provide personalized travel recommendations, itineraries, and real-time assistance. zIA adopts the persona of a knowledgeable Italian "auntie" to offer a friendly, engaging user experience in both Italian and English, supporting text and voice interactions. The architecture comprises a front-end web application, a backend for data processing and API management, and cloud infrastructure for deployment. The system incorporates various data sources, including local tourism databases, third-party websites, and real-time event information. The chatbot uses a two-step conversational approach to gather user preferences and generate tailored travel plans. Future enhancements include benchmarking commercial LLMs against open-source alternatives, exploring small language models (SLMs) orchestrated by an agent-based engine called MindStream, migrating to a mobile app, and integrating text-to-speech and AR technologies. The authors highlight the potential advantages of SLMs for personalization, sustainability, edge computing compatibility, and enhanced data protection. The paper concludes by discussing the benefits of GenAI chatbots in tourism, including deep personalization, natural conversations, real-time support, and continuous learning capabilities.

- Mubarak et al. (2023) proposes Att-KGCN, an improved Attention Knowledge Graph Convolution Network model for recommending tourist attractions. The model leverages a knowledge graph of tourist attractions and user interaction data to make personalized recommendations. It employs a graph convolutional network to represent entities as vectors through neighborhood aggregation, and introduces an attention mechanism to calculate weights between target attractions and adjacent entities. The model was evaluated on a dataset of Socotra Island tourism information, containing 1,500 scenic spots, 2,229 tourists, and 6,091 score records. Experiments compared different aggregation functions (sum, concat, neighbor) and hyperparameters, with the sum aggregator performing best. The Att-KGCN model outperformed the baseline KGCN model, achieving AUC and F1-scores of 0.98 and 0.95 respectively, compared to 0.96 and 0.92 for KGCN. The authors found optimal performance with 8 neighbor entities sampled, an embedding dimension of 16, and a receptive field depth of 2. The model demonstrates improved effectiveness in recommending scenic spots by capturing higher-order structural and semantic information from the knowledge graph, while addressing the challenge of selecting suitable attraction attributes in the tourism domain.

- Chen et al. (2023a) introduces a novel non-stationary multi-armed bandit (MAB) framework with an auto-regressive (AR) reward structure to capture temporal dependencies in rapidly changing environments. The authors propose the AR2 algorithm, which employs alternating exploration-exploitation and restarting mechanisms to address challenges in balancing exploration-exploitation and managing outdated information. The algorithm's performance is evaluated using a robust dynamic regret metric. Theoretical analysis shows that AR2 achieves a regret upper bound of $O(c_2^0\alpha^2\sigma^2k^3\log(c_0\alpha\sigma\sqrt{k}))$ for $\alpha \in [0.5, 1)$, nearly matching the derived lower bound of $\Omega(k\alpha^2\sigma^2)$. The paper also presents an extension of AR2 for general AR-p processes and demonstrates its efficacy through a real-world case study on tourism demand prediction. Numerical experiments further validate

AR2's superiority over various benchmark algorithms in both stationary and non-stationary settings. The authors also explore relaxing the assumption of known AR parameters through maximum likelihood estimation, showing that AR2 maintains competitive performance even with estimated parameters. This work contributes significantly to the understanding of non-stationary bandits with temporal structures, offering insights applicable to recommendation systems, online advertising, and financial portfolio management.

- Ho et al. (2023b) presents SBTREC, a novel BERT-based Transformer framework for personalized tour recommendation that incorporates sentiment analysis. The algorithm analyzes users' check-ins, uploaded photos, and reviews to understand POI visit patterns and user preferences. SBTREC utilizes a BERT embedding model with a custom NEXTPOP gate to enhance POI prediction, considering factors like location, time constraints, and individual preferences. The method employs sentiment analysis on user reviews using S-BERT embeddings to refine recommendations. The algorithm was evaluated against nine baseline methods using datasets from eight cities, demonstrating superior performance with an average F1 score of 64.30%. SBTREC outperformed the next best algorithm (PPOIBERT) by 7.85 percentage points in F1 score. The framework's key innovations include the integration of sentiment analysis, the NEXTPOP gate for incorporating external factors like user comments and photo counts, and the ability to adapt to different scenarios without modification. Experiments showed consistent performance across diverse cities and POI themes, with improvements of up to 12.93% in F1 score for some datasets compared to baseline methods. The authors suggest future work could involve integrating additional information sources and conducting ablation studies to further validate the results.

- Ho et al. (2023a) introduces BTREC (BERT-based Trajectory Recommendation), an innovative algorithm for personalized tour itinerary recommendations. BTREC extends the POIBERT embedding algorithm by incorporating users' demographic information and past POI visits into a modified BERT language model. The algorithm treats users' trajectories as sentences and POI visits as words, employing a Transformer architecture to capture contextual relationships. BTREC iteratively predicts the next POI using a Masked Language Model approach, considering factors such as location, travel time, and individual preferences. The algorithm was evaluated against eight baseline methods using datasets from eight cities, demonstrating superior performance with an average F1 score of 63.55%, outperforming the next best algorithm (POIBERT) by 1.23 percentage points. BTREC showed consistent performance across diverse cities without requiring modifications, with improvements of up to 6.48% in F1 score for some datasets compared to previous implementations. The authors highlight the algorithm's ability to adapt to different scenarios and suggest potential future work in fine-tuning personalized embeddings for users with missing demographic information to further enhance recommendation accuracy.

- Ho and Lim (2022) introduce POIBERT, a novel algorithm for personalized tour itinerary recommendation using a BERT-based language model adapted to process sequences of Points of Interest (POIs). The authors model the tour recommendation problem as a sequential recommendation task, treating POIs as words and user trajectories as sentences. POIBERT employs an iterative approach to generate consecutive POIs, optimizing for time constraints and user preferences based on historical data. The algorithm uses bootstrapping to estimate visit durations with confidence intervals, enhancing the accuracy of time-sensitive predictions. Evaluated on a Flickr dataset covering seven cities, POIBERT outperforms eight baseline sequence prediction algorithms, achieving an average F1 score of up to 69.85% (in Delhi), with improvements of up to 19.99 percentage points over the next best method. The authors also propose POILSTM, an LSTM-based variant, which shows comparable performance. POIBERT demonstrates adaptability across different cities without modification, suggesting its potential for broader application in tour recommendation systems. The paper concludes

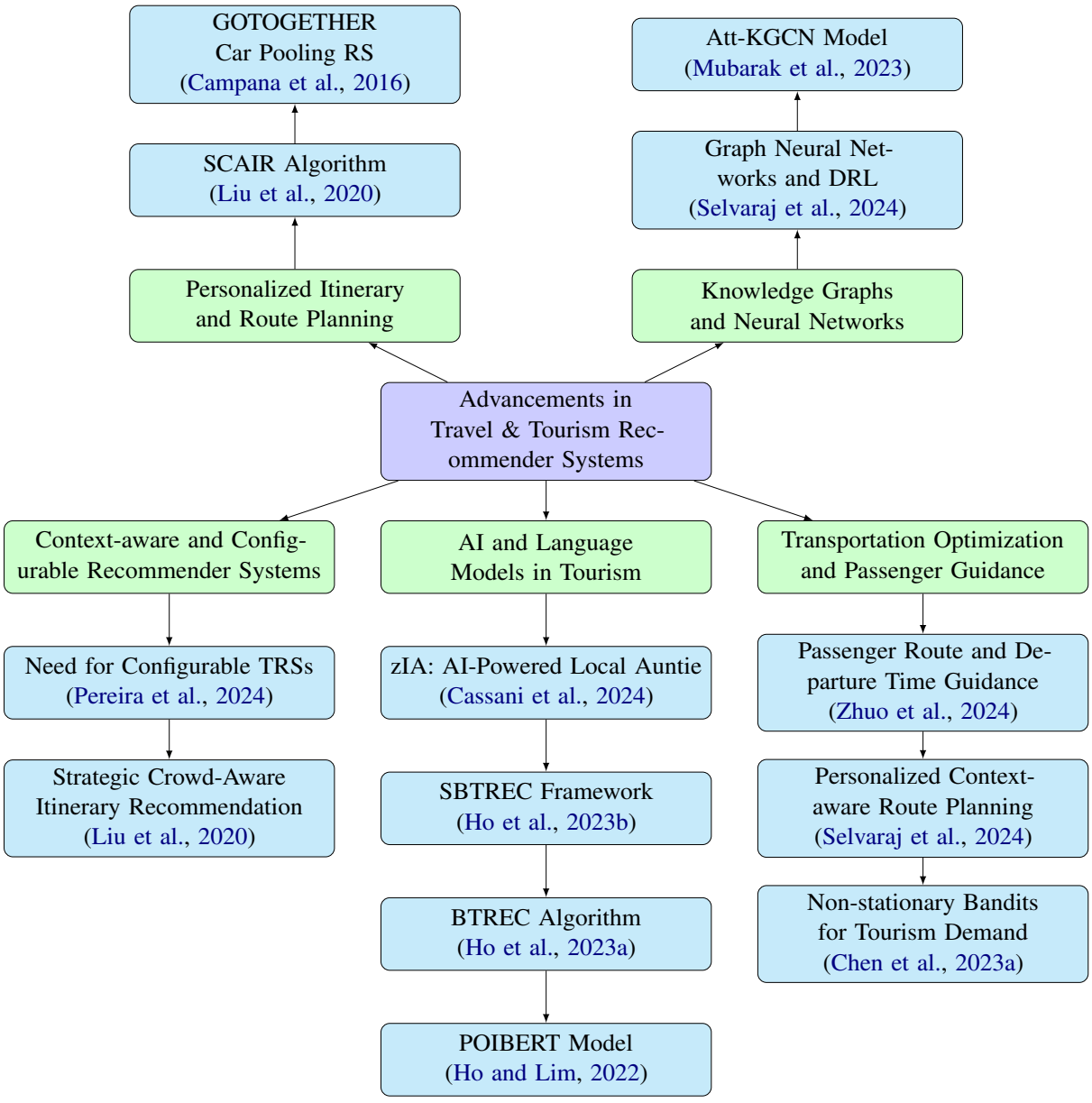

Figure 8: Concept map illustrating the advancements in Travel & Tourism recommender systems. The map is divided into five primary categories: **Context-aware and Configurable Recommender Systems**, **AI and Language Models in Tourism**, **Transportation Optimization and Passenger Guidance**, **Personalized Itinerary and Route Planning**, and **Knowledge Graphs and Neural Networks**. Under each category, key research studies and methods are outlined, showing the progression and relationships between different approaches.

by suggesting future work on adapting other language models and creating a HuggingFace interface for POIBERT.

- Liu et al. (2020) introduces the Strategic and Crowd-Aware Itinerary Recommendation (SCAIR) algorithm to address the Selfish Routing problem in itinerary planning, where individual optimization leads to suboptimal social welfare. The authors formulate the problem as a strategic game and model it using finite Markov chains to optimize for multiple travelers simultaneously. SCAIR considers crowd behavior and dynamically updates queuing times based on expected visitor numbers. The algorithm is evaluated against three baselines (Distance Optimization, Popularity Optimization, and Popularity over Distance Optimization) using a theme park dataset from Epcot and Disney Hollywood Studios. Experiments vary arrival intervals and simulation times, measuring average popularity, queuing time, and utility. Results show SCAIR significantly outperforms baselines, reducing queuing time ratios by 78.9% to 93.4% while maintaining comparable popularity scores and consistently higher utility across different time budgets. The authors prove the NP-hardness of both the path-finding and social welfare optimization problems, and demonstrate SCAIR's effectiveness in balancing individual preferences with overall system efficiency in crowd-aware itinerary recommendation scenarios. CopyRetryClaude can make mistakes. Please double-check responses.

- Campana et al. (2016) present GOTOGETHER, a personalized car pooling recommender system that utilizes learning-to-rank techniques to dynamically infer individual user preferences. The system employs an online, pairwise learning algorithm that leverages users' ride selection history to predict their personal ranking models. GOTOGETHER uses a weighted linear combination of ride features, including walking distances, pickup delays, and social similarity between drivers and passengers. The algorithm balances exploration and exploitation using an ϵ-greedy strategy. Evaluation was conducted using a dataset generated from Twitter and Foursquare data, simulating plausible mobility patterns in New York City. Experiments in static and dynamic scenarios demonstrated the algorithm's ability to quickly learn and adapt to user preferences, with exploration rates of 0.1-0.2 providing optimal performance. The system achieved high ranking accuracy for best ride matches and improved success probabilities compared to random selection. Performance degraded with higher acceptance thresholds due to fewer learning examples. A mobile application was developed and deployed in a corporate setting for real-world testing. The authors suggest future work including multi-modal transportation scenarios and incorporating additional data sources and ride features.

3.7 Education

The landscape of educational technology is undergoing a profound transformation, driven by the integration of advanced artificial intelligence, particularly Large Language Models (LLMs), into recommender systems and learning platforms. Recent innovations are addressing critical challenges in personalized learning, student engagement, and course recommendations across various educational contexts, from MOOCs to traditional university settings.

Frameworks like Tailor-Mind and RAMO are leveraging LLMs to enhance self-regulated learning and tackle the "cold start" problem in course recommendations, respectively. These advancements are complemented by novel approaches in collaborative filtering, such as the SimCE loss function, which is improving the accuracy and efficiency of recommendation algorithms.

Personalization is reaching new heights with systems like the MBTI-based gamification model and the GenRec framework, which are tailoring educational experiences to individual learning styles and preferences. Concurrently, there's a growing emphasis on explainability and fairness in educational AI, exemplified by models like UPGPR for MOOCs and the Exposure-Aware Cascading Bandit algorithm.

These developments are not only enhancing the quality of recommendations but also transforming how students interact with educational content, fostering deeper engagement, and improving learning outcomes. As the field progresses, it promises to deliver more adaptive, engaging, and effective educational experiences, while also addressing crucial

ethical considerations in AI-driven education.

- Gao et al. (2024) introduce a conceptual framework for integrating fine-tuned Large Language Models into interactive visualization systems for domain-specific tasks. The framework addresses three key alignments: domain problems with LLMs, visualizations with LLMs, and interactions with LLMs. The authors apply this framework to the educational domain, developing Tailor-Mind, an intelligent visualization system for self-regulated learning (SRL) in artificial intelligence. Tailor-Mind incorporates a fine-tuned LLM based on Baichuan2-7B-chat, trained on a custom dataset of 74,932 entries across four scenarios. The system features a chat interface, file preview, knowledge mindmap, question recommendations, and a learning path visualization to support the SRL process. Evaluation of the fine-tuned model shows superior performance compared to baseline models, with an average F1 score of 4.30 in human evaluations. A user study with 24 participants demonstrates Tailor-Mind's effectiveness in enhancing learning efficiency, depth of understanding, and engagement in SRL tasks compared to using GPT-4 alone. The paper concludes by discussing the framework's generalizability, potential enhancements, and limitations, providing valuable insights for developing intelligent, domain-specific visualization systems.

- Rao and Lin (2024) introduce RAMO (Retrieval-Augmented Generation for MOOCs), a novel course recommender system that leverages Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) to address the "cold start" problem in MOOCs. The system utilizes a Coursera dataset and employs GPT-3.5 Turbo as its backend model, integrated with the LangChain framework. RAMO outperforms traditional recommender systems in handling new users without historical data, providing relevant course suggestions based on general queries. The study demonstrates RAMO's flexibility in adapting to various user prompts and retrieval templates, enabling personalized course recommendations. Comparative analysis shows RAMO's superiority over non-LLM systems in response time and recommendation relevance, especially for new users. The RAG approach enhances the LLM's performance by incorporating external knowledge bases, resulting in more tailored and accurate recommendations. While the study presents promising results, it acknowledges limitations such as the lack of comprehensive evaluations and user studies. Future work aims to conduct thorough evaluations, enhance system performance, and deploy RAMO on a dedicated e-learning platform to gather real-world user data for further refinement and validation.

- Yang et al. (2024) introduce SimCE (Simplified Sampled Softmax Cross-Entropy Loss), a novel loss function for collaborative filtering in recommender systems. SimCE addresses limitations of existing methods like BPR (Bayesian Personalized Ranking) and SSM (Sampled Softmax Cross-Entropy) by simplifying SSM using its upper bound. The authors conduct extensive experiments on 12 benchmark datasets using Matrix Factorization and LightGCN backbones, demonstrating SimCE's superior performance over BPR and SSM in 93 out of 96 comparison instances, with improvements up to 68.72%. SimCE consistently benefits from multiple negative samples during training, showing optimal performance with 64 negative samples for large-scale datasets. The study also analyzes the impact of the margin hyperparameter γ, finding that larger values may enhance performance, particularly for Matrix Factorization. SimCE achieves comparable training efficiency to SSM while significantly outperforming BPR in convergence speed and accuracy. The authors provide PyTorch-style pseudo-code for easy implementation and integration of SimCE into existing frameworks. This work highlights the importance of loss function design in recommender systems and opens avenues for future research in areas such as knowledge graph recommendations and sequential recommendations.

- Ibisu (2024) present the development of a gamification model for personalized e-learning based on the Myers-Briggs Type Indicator (MBTI) cognitive core. The research employed a mixed-methods approach, com-

bining literature review, expert surveys, and system implementation. A taxonomy of game elements for educational environments was extended and mapped to MBTI cognitive cores based on expert recommendations. The resulting model was implemented as an asynchronous online course using WordPress and LearnDash. The system was evaluated by 37 participants over 3 months, assessing engagement (appeal, emotion, user-centricity, satisfaction) and educational usability (clarity, error correction, feedback). Results showed positive ratings across all criteria, with emotion scoring highest ($\mu = 4.5$) for engagement and feedback highest ($\mu = 4.8$) for educational usability. Sensing-Feeling learners consistently reported better results than other cognitive pairs. The study concludes that personalized gamification based on MBTI cognitive cores can enhance learner engagement and educational outcomes in e-learning environments, contributing to the field of Information Systems by providing a novel approach to tailoring educational experiences.

- Xu et al. (2024c) explore the integration of large language models (LLMs) with collaborative filtering (CF) algorithms to enhance e-commerce recommendation systems. The study proposes a framework that combines user behavior data with LLM-guided features to generate personalized recommendations. The methodology involves calculating user similarities, implementing both user-based and item-based CF algorithms, and leveraging LLMs to enrich feature representations and similarity calculations. Experiments compare traditional CF algorithms (UserCF and ItemCF) with their LLM-enhanced counterparts (T-UserCF and T-ItemCF) using Mean Absolute Error (MAE) as the primary evaluation metric. Results demonstrate that LLM-enhanced algorithms consistently outperform traditional methods across various numbers of neighbors (K), with T-UserCF achieving optimal performance at K=10. The integration of LLMs improves recommendation accuracy and personalization by enabling more nuanced understanding of user preferences and item characteristics. The study concludes that the synergy between LLMs and CF algorithms offers significant potential for advancing e-commerce recommendation systems, addressing limitations of traditional approaches and delivering enhanced user experiences.

- Frej et al. (2024) present an explainable recommendation system for Massive Open Online Courses (MOOCs) using reinforcement learning and graph reasoning over knowledge graphs. The authors propose Unrestricted Policy-Guided Path Reasoning (UPGPR), which extends the Policy-Guided Path Reasoning (PGPR) approach to generate interpretable recommendations without relying on predefined path patterns. UPGPR is evaluated on two public MOOC datasets (COCO and Xuetang) and outperforms baseline models in ranking metrics. The study includes a user evaluation with 25 participants to assess preferences for path-based explanations. Results show that learners prefer path-based explanations over popularity-based ones and value explanations using course categories and teachers over those based on other learners. The study reveals a trade-off between path length and interpretability, with longer paths improving model performance but decreasing user comprehension. Participants in the "Learn" condition (intrinsic motivation) desired more detailed explanations compared to those in the "Credits" condition (extrinsic motivation). The research contributes to the field of explainable AI in education by demonstrating the efficacy of path-based recommendations and highlighting the importance of balancing accuracy with interpretability in educational recommendation systems.

- Mansoury et al. (2024) addresses exposure bias in online learning-to-rank recommendation systems, specifically focusing on Linear Cascading Bandits. The authors propose an Exposure-Aware Cascading Bandit (EACB) algorithm that incorporates a novel reward model considering both user feedback and item position in the recommendation list. The EACB algorithm adjusts rewards and penalties for clicked and unclicked items based on their position, aiming to mitigate exposure bias while maintaining recommendation accuracy. Experiments on MovieLens and Yahoo Music datasets demonstrate that EACB outperforms baseline methods, in-

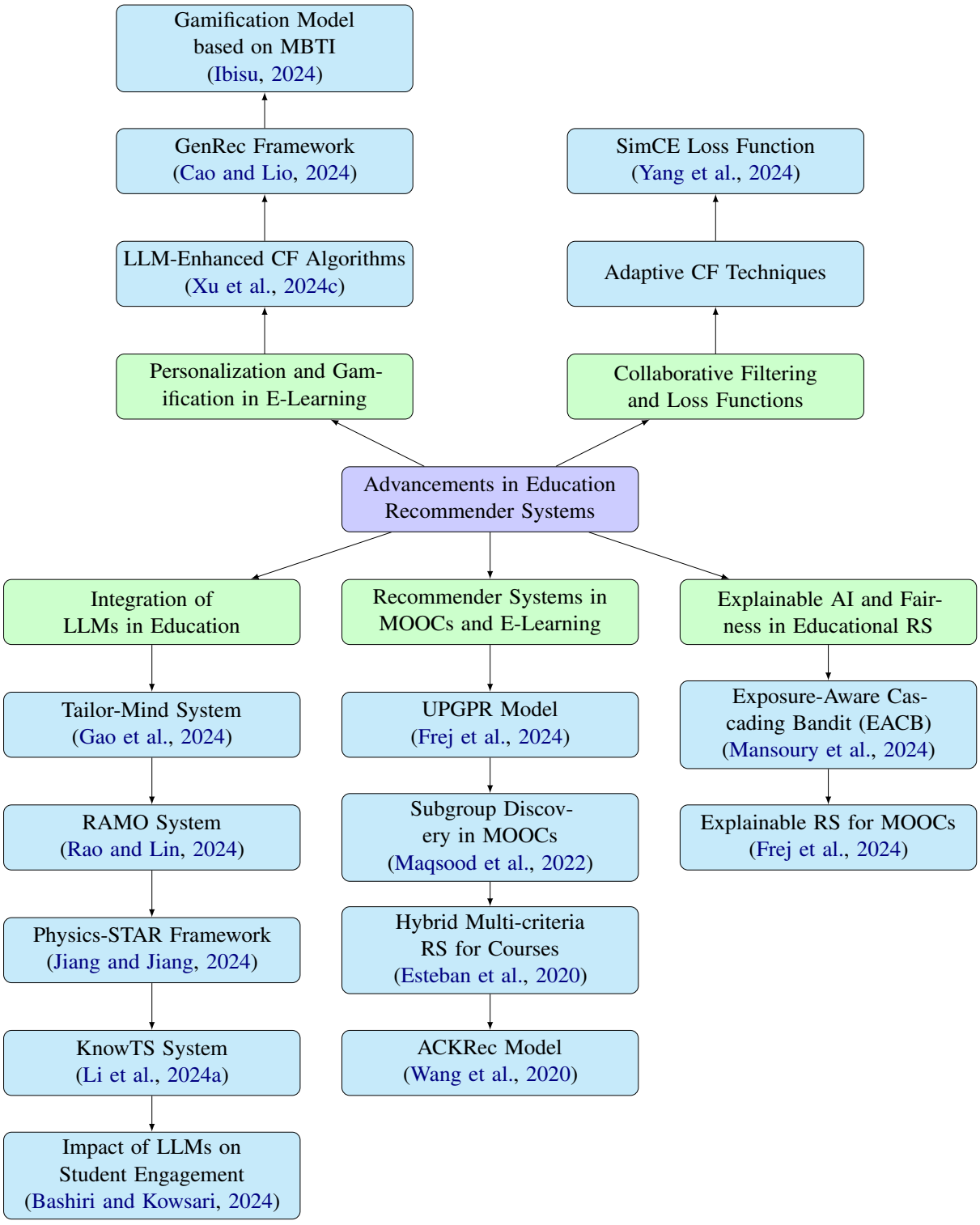

Figure 9: Concept map illustrating the advancements in education recommender systems. The map is divided into five primary categories: **Integration of LLMs in Education**, **Recommender Systems in MOOCs and E-Learning**, **Personalization and Gamification in E-Learning**, **Collaborative Filtering and Loss Functions**, and **Explainable AI and Fairness in Educational RS**. Under each category, key research studies and methods are outlined, showing the progression and relationships between different approaches. Cross-connections indicate how certain methodologies influence or relate to others across categories, highlighting the interdisciplinary nature of recent advancements in education recommender systems.

cluding LinUCB, EARSLinUCB, and FRM-LinUCB, across multiple exposure fairness metrics. The authors provide a theoretical analysis, proving a high-probability upper regret bound for EACB that matches the bound of the original cascading bandits. Sensitivity analyses reveal the impact of exploration parameters and penalization coefficients on the algorithm's performance. The study contributes to addressing long-term exposure bias in online recommendation systems, offering both empirical and theoretical evidence for the effectiveness of the proposed approach in balancing item exposure while preserving recommendation quality.

- Li et al. (2024a) introduce KnowTS, a novel knowledge tagging system for math questions utilizing Large Language Models (LLMs). The system outperforms traditional machine learning approaches, especially in scenarios with limited or no annotated data. KnowTS incorporates a zero-shot inference pipeline and a few-shot learning capability, leveraging LLMs' strong in-context learning abilities. To optimize demonstration selection, the authors propose FlexSDR (Flexible Sequential Demonstration Retriever), an RL-based algorithm that dynamically selects and adapts the number of demonstrations for each query. FlexSDR introduces an early stop mechanism and intermediate rewards, improving performance while using fewer demonstrations compared to existing methods. Experiments on the MathKnowCT dataset, covering 24 math concepts for grades 1-3, demonstrate KnowTS's superior performance over baselines like embedding similarity and pre-trained language model fine-tuning. The system achieves up to 85% F1 score in zero-shot settings and shows significant improvements with few-shot learning. Ablation studies confirm the effectiveness of FlexSDR's components, including the early stop mechanism and step-wise reward design. The research contributes to advancing automated knowledge tagging in educational contexts, offering a promising approach for handling complex mathematical concepts with minimal annotated data.

- Jiang and Jiang (2024) introduce Physics-STAR, a novel framework for a Large Language Model (LLM)-powered tutoring system designed to enhance personalized learning in high school physics education. The system utilizes a three-step approach: knowledge explanation, error analysis, and review suggestion, structured using the Situation-Task-Action-Result (STAR) methodology. An experimental study with 12 high school sophomores compared Physics-STAR against traditional teacher-led lectures and generic LLM tutoring. Results demonstrated that Physics-STAR significantly improved students' performance and efficiency, particularly in complex information-based problems, where average scores increased by 100% and efficiency by 5.95%. The system excelled in facilitating deep learning and precise understanding of physics concepts through step-by-step guidance and reflective learning. The study emphasizes the importance of student-centered AI in education, highlighting the need for LLM-powered systems to prioritize error analysis, reflective learning, and the mastery of fundamental skills. The authors conclude that as LLM technology advances, multimodal knowledge interactions will further enhance personalized learning experiences in physics education and potentially other disciplines.

- Bashiri and Kowsari (2024) explores the transformative impact of Large Language Models (LLMs) and AI tools on student engagement in social media and educational platforms. Drawing from UniversityCube data, the study reveals that AI-enhanced social media platforms lead to improved academic performance, critical thinking skills, and collaborative project engagement among students. The integration of LLMs facilitates personalized learning experiences, real-time feedback, and efficient peer-to-peer communication. The paper also examines the role of AI-powered image generators in enhancing visual communication of complex concepts in academia. Time series analysis of student usage patterns in entertainment and educational social networks from 2011 to 2024 shows a declining trend in entertainment network usage and an increasing trend in educational network usage. The study employs machine learning techniques, including exploratory data analysis, feature engineering, and clustering, to

uncover deeper insights into usage patterns. The authors conclude that the combination of educational social networks with LLMs and AI-driven visualization tools offers a multifaceted approach to learning, fostering personalized experiences, collaboration, and improved comprehension of complex subjects. The findings underscore the potential of AI-driven tools to create enriched, efficient, and supportive educational environments in the digital age.

- Cao and Lio (2024) introduce GenRec, a novel generative framework for personalized sequential recommendation. GenRec utilizes an encoder-decoder Transformer architecture and formulates sequential recommendation as a sequence-to-sequence generation task. The model employs a cloze task for pretraining and finetuning, learning bidirectional sequential patterns by masking random items in user interaction sequences. Unlike previous approaches, GenRec does not rely on manually designed prompts or additional large pretraining corpora. The model incorporates cross-modal embeddings including token, positional, user ID, and item ID embeddings to capture personalized information. Experimental results on Amazon Sports, Amazon Beauty, and Yelp datasets demonstrate that GenRec outperforms or matches state-of-the-art baselines across multiple metrics, including Hit Ratio and Normalized Discounted Cumulative Gain. An ablation study confirms the effectiveness of the pretraining objective. The authors highlight GenRec's efficiency, requiring only a few hours of training in low-resource settings, and its potential for democratizing large language models in the sequential recommendation domain.

- Maqsood et al. (2022) propose a novel subgroup discovery (SD) approach using MapReduce for categorizing and describing different types of learners in Massive Open Online Courses (MOOCs). The methodology extends the FP-Growth algorithm to handle large datasets and incorporates a threshold for the number of courses each discovered rule should satisfy. A post-processing step removes redundant subgroups. The approach is evaluated using de-identified data from 16 MITx and HarvardX courses on the edX platform, comprising 641,138 student records. The MapReduce implementation outperforms traditional sequential SD approaches, achieving near-constant runtime across different courses. The algorithm discovers interpretable IF-THEN rules describing four learner categories: Certified, Only Explored, Only Viewed, and Only Registered. The post-processing step significantly reduces the number of discovered subgroups, enhancing interpretability. The study finds distinct patterns for each learner category, such as high grades and engagement for Certified learners, medium engagement for Only Explored learners, and low engagement for Only Viewed learners. No consistent patterns were found for Only Registered learners across all courses. The discovered rules can be used for learner description, prediction, recommendation, and personalization in MOOCs. The approach demonstrates scalability and effectiveness in mining meaningful patterns from large-scale MOOC data.

- Esteban et al. (2020) present a hybrid multi-criteria recommendation system for university elective courses, combining collaborative filtering (CF) and content-based filtering (CBF) approaches. The system utilizes multiple criteria related to both student information (ratings, grades, branch) and course information (professors, competences, knowledge area, contents). A genetic algorithm (GA) is employed to automatically optimize the system configuration, including weights for each criterion, similarity measures, and neighborhood size. The study uses real data from 95 Computer Science students at the University of Cordoba, Spain, covering 2500 ratings across 63 courses. Experimental results demonstrate that the hybrid approach outperforms individual CF and CBF models, achieving lower RMSE (0.971) and higher nDCG (0.682) scores. The GA-optimized configuration reveals that student ratings and professor information are the most influential criteria. The system is compared to existing approaches and shows superior performance in terms of accuracy and recommendation relevance. A case study further illustrates the system's ability to provide personalized

course recommendations. The authors conclude that considering multiple weighted criteria and combining CF with CBF leads to improved recommendation quality, highlighting the importance of proper parameter tuning in recommendation systems for educational contexts.

- Wang et al. (2020) introduce ACKRec, an attentional graph convolutional network approach for knowledge concept recommendation in Massive Open Online Courses (MOOCs). The authors model the MOOCs data as a heterogeneous information network (HIN) to capture complex relationships between users, courses, videos, teachers, and knowledge concepts. ACKRec leverages meta-path guided graph convolutional networks to learn entity representations, incorporating both content and context information. An attention mechanism adaptively fuses representations from different meta-paths to capture diverse user interests. The model is optimized using an extended matrix factorization approach. Experiments on real-world data from XuetangX demonstrate ACKRec's superior performance over state-of-the-art baselines across multiple metrics, including Hit Ratio, NDCG, MRR, and AUC. The authors conduct detailed analyses of meta-path combinations, model parameters, and provide a case study illustrating the effectiveness of the approach. ACKRec addresses the overlooked problem of knowledge concept recommendation in MOOCs, offering a more granular and personalized recommendation approach compared to course-level recommendations.

3.8 Bandits

Multi-Armed Bandits (MAB) represent an advanced approach to online testing, offering a more dynamic alternative to traditional A/B testing. This method uses machine learning algorithms to dynamically allocate traffic to better-performing variations while reducing exposure to underperforming ones, effectively balancing exploration of different options with exploitation of known high performers. MAB algorithms, such as epsilon-greedy, upper confidence bound, and Thompson sampling, allow for continuous optimization and faster identification of winning versions. They are particularly useful for short-term campaigns, scenarios with long-term dynamic changes, targeted user experiences, and large-scale automated optimization. While computationally more complex than standard A/B tests, MAB techniques can significantly reduce the opportunity cost of showing suboptimal experiences during testing periods. Some industry platforms now offer MAB solutions for various objectives, including accelerating the discovery of statistically significant variations and maximizing rewards while minimizing regret. This approach is especially valuable in recommender systems, where user preferences can change rapidly and multiple strategies may need to be tested simultaneously.

(Mabry et al., 2021) Bain & Company discusses the advantages of using multi-armed bandit algorithms in marketing over traditional A/B testing and multivariate testing. MABs dynamically optimize marketing messages, reducing the cost of testing due to lost conversions. They are particularly useful for frequent customer interactions and can maintain effectiveness over time, unlike traditional methods which may see declining results. MABs balance exploration of new options with exploitation of known successful strategies, allowing marketers to "earn while learning." The article highlights the use of contextual bandits, which incorporate user information to personalize experiences. MABs are especially valuable for online services, pricing experiments, and personalized marketing. To implement MABs effectively, companies need robust experimentation programs, accurate customer data, and product-focused teams. The article suggests that as first-party data becomes more crucial due to the decline of third-party cookies, MABs and AI-driven personalization will become increasingly important for marketers to effectively leverage their customer data.

Meta's (Facebook Research, 2021) research addresses the challenge of optimally allocating an advertiser's budget across multiple platforms (e.g., Facebook, Instagram) when both demand and value are unknown. The problem is modeled as a stochastic bandit problem with budgets, where an algorithm must decide on bids for different platforms to maximize advertiser value within a given budget and time horizon. The researchers propose a modified algorithm that performs well when the number of platforms is large and the total possible value is

small relative to the total number of auction participations, which is typical in online advertising. The algorithm uses a primal-dual approach and solves an optimization program at each time step without requiring a rounding step, leading to optimal regret guarantee. Tested on logged data, the algorithm shows superior performance compared to prior works and industry heuristics. This research has potential benefits for advertisers, users, and platforms by providing scalable, near-optimal bidding solutions for automated advertising products, and it bridges the gap between theoretical research and practical application in the field of "Bandits with Budgets.

Amazon researchers (Stein and Moerchen, 2020) have developed a flexible framework for solving multi-armed bandit problems, which are used to optimize decision-making in various scenarios. This framework models interactions as ordering finite lists of actions, with each action represented as a vector that can include contextual information. The system learns from feedback on chosen actions, balancing exploration of new options with exploitation of known good choices. The researchers demonstrated the framework's effectiveness in two applications presented at CIKM 2020: improving music recommendation rankings and natural language understanding for voice assistants. In testing, the approach led to significant improvements in user engagement with music recommendations and in correctly interpreting ambiguous voice commands. The framework's flexibility allows it to be adapted for various algorithms and applications, making it a powerful tool for enhancing customer experiences across different Amazon services.

Stitch Fix (Amadio, 2020) have integrated Multi-Armed Bandits into their existing experimentation platform, complementing traditional A/B testing methods. This enhancement leverages their flexible system of configuration parameters and randomization units to support a wide range of experiments across the business. The MAB implementation dynamically allocates traffic to better-performing variants, potentially reducing opportunity costs compared to static A/B tests. The platform supports standard policies like e-greedy and Thompson Sampling, as well as custom policies, with each MAB experiment utilizing a dedicated microservice for reward estimates. Data scientists can implement and update their own reward models, while the platform provides tools for automating microservice setup and data updates. The system also accommodates contextual bandits, allowing for more sophisticated decision-making based on environmental factors. This implementation seamlessly integrates with existing engineering applications, maintaining ease of use while offering advanced experimentation capabilities. Notably, Stitch Fix hints at a novel method for efficient, deterministic online computation of Thompson Sampling probabilities, promising further improvements to their MAB approach.

4 Evaluation

Evaluating the performance of recommendation models is crucial to ensure that they are effective and efficient in providing relevant recommendations to users. We summarize here some standard evaluation strategies and metrics that can be used.

4.1 Evaluation Strategies

The following are some common strategies used to evaluate the performance of a recommendation system:

- **A/B Testing:** This involves testing two versions of a recommendation system against each other to see which one performs better.
- **Offline Evaluation:** This involves evaluating the recommendation system using historical data, without deploying it to production.
- **Online Evaluation:** This involves evaluating the recommendation system in real-time, as it is being used by users.

These strategies can be used to evaluate the performance of a recommendation system and identify areas for improvement.

4.2 Evaluation Metrics

This section lists different metrics we can use to understand how accurate, diverse, and valuable the recommendations are. Evaluation metrics help us benchmark the performance of a recommendation system, improve the models, and measure its performance across these metrics.

4.2.1 Accuracy Metrics

Accuracy metrics are used to measure how well the recommendation system is performing in terms of predicting user preferences when compared to ground truth set from online user data or outsourced labels. These metrics include:

- **Precision:** The ratio of true positive predictions (correctly predicted relevant items) to the total number of predicted relevant items.
- **Recall:** The ratio of true positive predictions to the total number of actual relevant items.
- **F1 Score:** The harmonic mean of precision and recall, which gives equal weight to both metrics.
- **RMSE (Root Mean Squared Error):** A measure of the difference between predicted ratings and actual ratings, with lower values indicating better performance.
- **MAP (Mean Average Precision):** A measure of the average precision of the recommendation system across all users, with higher values indicating better performance.
- **Top-k Hit Ratio (H@K):** Measures the accuracy of a recommendation system by calculating the proportion of times the correct item appears in the top K recommendations. A higher H@K indicates better accuracy.
- **Normalized Discounted Cumulative Gain (N@K):** Evaluates the ranking quality by considering the positions of relevant items in the top K recommendations, with higher values reflecting better ranking performance.

4.2.2 Diversity Metrics

Diversity metrics are used to measure how diverse the recommendations are, and whether the system is recommending a wide range of items or just a narrow subset. These metrics include:

- **Repetitiveness:** A measure of how often the same items are recommended to the same user, with lower values indicating more diverse recommendations.
- **Coverage:** A measure of the percentage of items in the catalog that are being recommended, with higher values indicating more diverse recommendations.

4.2.3 Value Metrics

Value metrics are used to measure the downstream value of the recommendations. These metrics include, but are not limited to:

- **Purchases:** The number of items purchased by users after seeing them in their recommendations.
- **Friending:** The number of friend requests sent by users after seeing them in their recommendations.
- **Engagement:** The amount of time spent by users on the platform after seeing their recommendations, amount of shares, comments etc.
- **Healthcare Impact:** The number of patients or customers acquired through the recommendations.
- **Finance Outcomes:** The number of loans, investments, or other financial transactions generated through the recommendations.

4.2.4 Quality and Safety Metrics

We need metrics to measure the quality and safety of the recommendations, and ensure that they are not promoting harmful or inappropriate content. These metrics include:

- **Fairness:** A measure of whether the recommendations are biased towards certain groups of users or items.
- **Bias:** A measure of whether the recommendations are skewed towards certain types of content or topics.
- **Harm:** A measure of whether the recommendations are promoting harmful or inappropriate content, such as hate speech, nudity, etc.

4.2.5 Survey Metrics

Survey metrics are used to gather feedback from users about their satisfaction with the recommendations. These metrics include:

- **NPS (Net Promoter Score):** A measure of how likely users are to recommend the platform to others, with higher values indicating greater satisfaction.
- **Relevance:** A measure of how relevant the recommendations are to users' interests, with higher values indicating greater relevance.

- **Sentiment and Satisfaction:** A measure of users' overall feelings or attitudes towards the recommendations, including different emotions triggered and satisfaction levels with the recommendations.
- **Trust and Ease of Use:** A measure of how much users trust the recommendations and find them easy to use, including factors such as perceived accuracy, reliability, transparency, navigation, comprehension, and control.

5 Findings

- **Integration of advanced AI techniques:** The paper does highlight the increasing use of Large Language Models and Graph Neural Networks in recommender systems.
- **Domain-specific adaptations:** The review covers recommender systems across various domains including e-commerce, social media, entertainment, healthcare, and finance.
- **Multi-modal approaches:** There is mention of leveraging diverse data types, including text, images, and user behavior data.
- **Fairness and explainability:** The paper discusses the growing emphasis on fairness, especially in domains like healthcare.
- **Emerging applications:** New applications such as NFT recommendations are mentioned.
- **Temporal dynamics:** Some models incorporating time-aware features are discussed.
- **Privacy considerations:** There is some mention of privacy-preserving techniques, particularly in the healthcare domain.

6 Limitations

- Limited discussion of ethical concerns: The paper touches on fairness and bias in recommender systems, but there's limited in-depth discussion of ethical issues such as privacy concerns, filter bubbles, and the potential for recommender systems to reinforce societal biases.
- **Lack of long-term impact studies:** Most of the reviewed studies focus on short-term metrics like accuracy and engagement. There's a gap in understanding the long-term effects of recommender systems on user behavior, preferences, and overall well-being.
- **Generalizability across cultures:** The papers doesn't address how well these recommender systems perform across different cultural contexts. Recommendations that work well in one culture may not be as effective in others.
- **Computational resources:** The papers doesn't thoroughly discuss the computational requirements of the advanced AI techniques (like LLMs and GNNs) being used in recommender systems. This could be a significant limitation for widespread adoption, especially for smaller organizations.
- **Integration challenges:** While the papers discusses various novel approaches, it doesn't delve into the challenges of integrating these new techniques into existing systems and workflows.
- **User control and transparency:** There's limited discussion on how much control users have over the recommendations they receive and how transparent these systems are to end-users.
- **Robustness to adversarial attacks:** They don't address the vulnerability of recommender systems to adversarial attacks or manipulation, which is an important consideration in real-world applications.
- **Cold start problem:** While mentioned briefly, there could be more discussion on how well the latest techniques address the cold start problem for new users or items.
- **Interdisciplinary perspective:** The review focuses primarily on technical aspects. There's limited discussion from other relevant fields like psychology, sociology, or economics, which could provide valuable insights into user behavior and the broader impacts of recommender systems.
- **Reproducibility:** Papers doesn't discuss challenges related to reproducing the results of the studies, which is a common issue in recommender systems research due to the use of proprietary datasets and algorithms.

7 Conclusion

This comprehensive review of recent advancements in recommender systems across e-commerce, social media, entertainment, healthcare, and finance sectors reveals significant progress and diversification in the field. Several key trends have emerged, shaping the current landscape of recommender systems research and applications.

The integration of advanced AI techniques, particularly LLMs and GNNs, has shown promising results in enhancing recommendation accuracy and addressing longstanding challenges such as data sparsity and cold-start problems. Researchers are increasingly developing domain-specific adaptations, recognizing the unique requirements of different sectors and tailoring solutions accordingly.

Multi-modal and contextual approaches have gained traction, leveraging diverse data types (text, images, user behavior) and contextual information to provide more personalized and accurate recommendations. Concurrently, there is a growing emphasis on fairness, privacy, and explainability in recommender systems, especially in sensitive domains like healthcare and finance.

Efficiency and scalability remain crucial concerns as recommender systems are deployed at increasingly large scales. Novel evaluation metrics are being developed to move beyond traditional accuracy measures, considering factors such as diversity, serendipity, and long-term user satisfaction.

The expanding scope of recommender systems is evident in emerging applications, including NFT recommendations and personalized healthcare interventions. This diversification underscores the versatility and potential impact of recommender systems across various domains.

Looking ahead, several promising research directions emerge. Further exploration of LLMs in recommendation tasks, development of more robust and generalizable models, and addressing ethical concerns associated with recommender systems are likely to be key areas of focus. Additionally, enhancing real-time and streaming recommendations and improving cross-domain recommendation capabilities present exciting opportunities for innovation.

In conclusion, while significant advancements have been made in recommender systems research, there remain ample opportunities for further innovation. As the field continues to evolve, addressing domain-specific challenges, enhancing user experience, and ensuring ethical and responsible deployment of these technologies will be crucial in realizing the full potential of recommender systems across diverse applications.

References

Nimesh Agrawal, Anuj Kumar Sirohi, Jayadeva, and Sandeep Kumar. 2023. No prejudice! fair federated graph neural networks for personalized recommendation. *Preprint*, arXiv:2312.10080.

Brian Amadio. 2020. Multi-armed bandits and the stitch fix experimentation platform. Stitch Fix Technology.

Sahar Arshad, Seemab Latif, Ahmad Salman, and Saadia Irfan. 2023. Increasing profitability and confidence by using interpretable model for investment decisions. *Preprint*, arXiv:2312.16223.

Masoud Bashiri and Kamran Kowsari. 2024. Transformative influence of llm and ai tools in student social media engagement: Analyzing personalization, communication efficiency, and collaborative learning. *Preprint*, arXiv:2407.15012.

Oliver Baumann, Durgesh Nandini, Anderson Rossanez, Mirco Schoenfeld, and Julio Cesar dos Reis. 2024. How to surprisingly consider recommendations? a knowledge-graph-based approach relying on complex network metrics. *Preprint*, arXiv:2405.08465.

Lex Beattie, Isabel Corpus, Lucy H. Lin, and Praveen Ravichandran. 2023. Evaluation framework for understanding sensitive attribute association bias in latent factor recommendation algorithms. *Preprint*, arXiv:2310.20061.

Nikita Bhalla, Adam Lechowicz, and Cameron Musco. 2023. Local edge dynamics and opinion polarization. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23. ACM.

Akanksha Bindal, Sudarshan Ramanujam, Dave Golland, TJ Hazen, Tina Jiang, Fengyu Zhang, and Peng Yan. 2024. Lipost: Improved content understanding with effective use of multi-task contrastive learning. *Preprint*, arXiv:2405.11344.

Mattia Giovanni Campana, Franca Delmastro, and Raffaele Bruno. 2016. A machine-learned ranking algorithm for dynamic and personalised car pooling services. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE.

Panfeng Cao and Pietro Lio. 2024. Genrec: Generative personalized sequential recommendation. *Preprint*, arXiv:2407.21191.

Alexio Cassani, Michele Ruberl, Antonio Salis, Giacomo Giannese, and Gianluca Boanelli. 2024. zia: a genai-powered local auntie assists tourists in italy. *Preprint*, arXiv:2407.11830.

Sarah H. Cen, Andrew Ilyas, and Aleksander Madry. 2023. User strategization and trustworthy algorithms. *Preprint*, arXiv:2312.17666.

Jianguo Chen, Kenli Li, Zhuo Tang, Kashif Bilal, and Keqin Li. 2016. A parallel patient treatment time prediction algorithm and its applications in hospital queuing-recommendation in a big data environment. *IEEE Access*, 4:1767–1783.

Qinyi Chen, Negin Golrezaei, and Djallel Bouneffouf. 2023a. Non-stationary bandits with auto-regressive temporal dependency. *Preprint*, arXiv:2210.16386.

Weixin Chen, Li Chen, Yongxin Ni, Yuhan Zhao, Fajie Yuan, and Yongfeng Zhang. 2023b. Fmmrec: Fairness-aware multimodal recommendation. *Preprint*, arXiv:2310.17373.

Minjoo Choi, Seonmi Kim, Yejin Kim, Youngbin Lee, Joohwan Hong, and Yongjae Lee. 2024. A recommender system for nft collectibles with item feature. *Preprint*, arXiv:2403.18305.

Munki Chung, Yongjae Lee, and Woo Chang Kim. 2023. Mean-variance efficient collaborative filtering for stock recommendation. *Preprint*, arXiv:2306.06590.

Nathan Corecco, Giorgio Piatti, Luca A. Lanzendörfer, Flint Xiaofeng Fan, and Roger Wattenhofer. 2024. An llm-based recommender system environment. *Preprint*, arXiv:2406.01631.

Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations.

Sunhao Dai, Changle Qu, Sirui Chen, Xiao Zhang, and Jun Xu. 2024. Recode: Modeling repeat consumption with neural ode. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, volume 31 of *SIGIR 2024*, page 2599–2603. ACM.

Kayla Duskin, Joseph S. Schafer, Jevin D. West, and Emma S. Spiro. 2024. Echo chambers in the age of algorithms: An audit of twitter's friend recommender system. *Preprint*, arXiv:2404.06422.

Juan Pablo Equihua, Maged Ali, Henrik Nordmark, and Berthold Lausen. 2023. Sequence-aware item recommendations for multiply repeated user-item interactions. *Preprint*, arXiv:2304.00578.

A. Esteban, A. Zafra, and C. Romero. 2020. Helping university students to choose elective courses by using a hybrid multi-criteria recommendation system with genetic optimization. *Knowledge-Based Systems*, 194:105385.

Facebook Research. 2021. Auto-placement of ad campaigns using multi-armed bandits. Facebook Research.

Matteo Ferrante, Matteo Ciferri, and Nicola Toschi. 2024. R&b – rhythm and brain: Cross-subject decoding of music from human brain activity. *Preprint*, arXiv:2406.15537.

Jibril Frej, Neel Shah, Marta Knezevic, Tanya Nazaretsky, and Tanja Käser. 2024. Finding paths for explainable mooc recommendation: A learner perspective. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, LAK '24. ACM.

Lin Gao, Jing Lu, Zekai Shao, Ziyue Lin, Shengbin Yue, Chiokit Ieong, Yi Sun, Rory James Zauner, Zhongyu Wei, and Siming Chen. 2024. Fine-tuned large language model for visualization system: A study on self-regulated learning in education. *Preprint*, arXiv:2407.20570.

Ashraf Ghiye, Baptiste Barreau, Laurent Carlier, and Michalis Vazirgiannis. 2023. Adaptive collaborative filtering with personalized time decay functions for financial product recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23. ACM.

Siyi Guo, Keith Burghardt, Valeria Pantè, and Kristina Lerman. 2024. Somer: Multi-view user representation learning for social media. *Preprint*, arXiv:2405.05275.

Lars Hillebrand, Armin Berger, Tobias Deußer, Tim Dilmaghani, Mohamed Khaled, Bernd Kliem, Rüdiger Loitz, Maren Pielka, David Leonhard, Christian Bauckhage, and Rafet Sifa. 2023. Improving zero-shot text matching for financial auditing with large language models. *Preprint*, arXiv:2308.06111.

Ngai Lam Ho, Roy Ka-Wei Lee, and Kwan Hui Lim. 2023a. Btrec: Bert-based trajectory recommendation for personalized tours. *Preprint*, arXiv:2310.19886.

Ngai Lam Ho, Roy Ka-Wei Lee, and Kwan Hui Lim. 2023b. Sbtrec- a transformer framework for personalized tour recommendation problem with sentiment analysis. *Preprint*, arXiv:2311.11071.

Ngai Lam Ho and Kwan Hui Lim. 2022. Poibert: A transformer-based model for the tour recommendation problem. *Preprint*, arXiv:2212.13900.

Xinyan Hu, Meena Jagadeesan, Michael I. Jordan, and Jacob Steinhardt. 2023. Incentivizing high-quality content in online recommender systems. *Preprint*, arXiv:2306.07479.

Ting-Ji Huang, Jia-Qi Yang, Chunxu Shen, Kai-Qi Liu, De-Chuan Zhan, and Han-Jia Ye. 2024. Improving llms for recommendation with out-of-vocabulary tokens. *Preprint*, arXiv:2406.08477.

Afvensu Enoch Ibisu. 2024. Development of a gamification model for personalized e-learning. *Preprint*, arXiv:2404.15301.

Guillermo Iglesias, Edgar Talavera, Jesús Troya, Alberto Díaz-Álvarez, and Miguel García-Remesal. 2024. Artificial intelligence model for tumoral clinical decision support systems. *Computer Methods and Programs in Biomedicine*, 253:108228.

Amit Kumar Jaiswal. 2024a. Towards a theoretical understanding of two-stage recommender systems. *Preprint*, arXiv:2403.00802.

Amit Kumar Jaiswal. 2024b. Towards a theoretical understanding of two-stage recommender systems. *Preprint*, arXiv:2403.00802.

Suvendu Jena, Jilei Yang, and Fangfang Tan. 2023. Unlocking sales growth: Account prioritization engine with explainable ai. *Preprint*, arXiv:2306.07464.

Zhoumingju Jiang and Mengjun Jiang. 2024. Beyond answers: Large language model-powered tutoring system in physics education for deep learning and precise understanding. *Preprint*, arXiv:2406.10934.

Erkang Jing, Yezheng Liu, Yidong Chai, Shuo Yu, Longshun Liu, Yuanchun Jiang, and Yang Wang. 2024. Personalized music recommendation with a heterogeneity-aware deep bayesian network. *Preprint*, arXiv:2406.14090.

Haesun Joung and Kyogu Lee. 2024. Music auto-tagging with robust music representation learned via domain adversarial training. *Preprint*, arXiv:2401.15323.

Chinmaya Kausik, Kevin Tan, and Ambuj Tewari. 2024. Leveraging offline data in linear latent bandits. *Preprint*, arXiv:2405.17324.

Jooeun Kim, Jinri Kim, Kwangeun Yeo, Eungi Kim, Kyoung-Woon On, Jonghwan Mun, and Joonseok Lee. 2024a. General item representation learning for cold-start content recommendations. *Preprint*, arXiv:2404.13808.

Sein Kim, Hongseok Kang, Seungyoon Choi, Donghyun Kim, Minchul Yang, and Chanyoung Park. 2024b. Large language models meet collaborative filtering: An efficient all-round llm-based recommender system. *Preprint*, arXiv:2404.11343.

Seonmi Kim, Youngbin Lee, Yejin Kim, Joohwan Hong, and Yongjae Lee. 2023. Nfts to mars: Multi-attention recommender system for nfts. *Preprint*, arXiv:2306.10053.

Seong Jin Lee, Will Wei Sun, and Yufeng Liu. 2024. Low-rank online dynamic assortment with dual contextual information. *Preprint*, arXiv:2404.17592.

Hang Li, Tianlong Xu, Jiliang Tang, and Qingsong Wen. 2024a. Knowledge tagging system on math questions via llms with flexible demonstration retriever. *Preprint*, arXiv:2406.13885.

Jun Li, Jingjian Wang, Hongwei Wang, Xing Deng, Jielong Chen, Bing Cao, Zekun Wang, Guanjie Xu, Ge Zhang, Feng Shi, and Hualei Liu. 2023. Fragment and integrate network (fin): A novel spatial-temporal modeling based on long sequential behavior for online food ordering click-through rate prediction. *Preprint*, arXiv:2308.15703.

Jundong Li, Jiliang Tang, Yilin Wang, Yali Wan, Yi Chang, and Huan Liu. 2018. Understanding and predicting delay in reciprocal relations. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, WWW '18. ACM Press.

Zongwei Li, Lianghao Xia, and Chao Huang. 2024b. Recdiff: Diffusion model for social recommendation. *Preprint*, arXiv:2406.01629.

Jan Malte Lichtenberg, Alexander Buchholz, and Pola Schwöbel. 2024. Large language models as recommender systems: A study of popularity bias. *Preprint*, arXiv:2406.01285.

Junhua Liu, Kristin L. Wood, and Kwan Hui Lim. 2020. Strategic and crowd-aware itinerary recommendation. *Preprint*, arXiv:1909.07775.

Kang Liu, Feng Xue, Dan Guo, Le Wu, Shujie Li, and Richang Hong. 2023. Megcf: Multimodal entity graph collaborative filtering for personalized recommendation. *ACM Transactions on Information Systems*, 41(2):1–27.

Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Zijian Zhang, Feng Tian, and Yefeng Zheng. 2024a. Large language model distilling medication recommendation model. *Preprint*, arXiv:2402.02803.

Sicen Liu, Xiaolong Wang, Xianbing Zhao, and Hao Chen. 2024b. Dkinet: Medication recommendation via domain knowledge informed deep learning. *Preprint*, arXiv:2305.19604.

Chenyang Lyu, Manh-Duy Nguyen, Van-Tu Ninh, Liting Zhou, Cathal Gurrin, and Jennifer Foster. 2023. Dialogue-to-video retrieval. *Preprint*, arXiv:2303.16761.

Joshua Mabry, Janani Sriram, and Richard Lichtenstein. 2021. Rolling out multiarmed bandits for fast, adaptive experimentation. Bain & Company.

Masoud Mansoury, Bamshad Mobasher, and Herke van Hoof. 2024. Mitigating exposure bias in online learning to rank recommendation: A novel reward model for cascading bandits. *Preprint*, arXiv:2408.04332.

Rabia Maqsood, Paolo Ceravolo, Cristóbal Romero, and Sebastián Ventura. 2022. Modeling and predicting students' engagement behaviors using mixture markov models. *Knowledge and Information Systems*, 64(5):1349–1384.

Lilian Marey, Bruno Sguerra, and Manuel Moussallam. 2024. Modeling activity-driven music listening with pace. *Preprint*, arXiv:2405.01417.

Thomas M. McDonald, Lucas Maystre, Mounia Lalmas, Daniel Russo, and Kamil Ciosek. 2023. Impatient bandits: Optimizing recommendations for the long-term without delay. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23. ACM.

M. Jeffrey Mei, Oliver Bembom, and Andreas F. Ehmann. 2024. Negative feedback for music personalization. *Preprint*, arXiv:2406.04488.

Gabriel Mendonça, Matheus Santos, André Gonçalves, and Yan Almeida. 2023. Rethinking financial service promotion with hybrid recommender systems at picpay. *Preprint*, arXiv:2310.10268.

Armin Moradi, Nicola Neophytou, and Golnoosh Farnadi. 2024. Advancing cultural inclusivity: Optimizing embedding spaces for balanced music recommendations. *Preprint*, arXiv:2405.17607.

Ahmad A. Mubarak, JingJing Li, and Han Cao. 2023. Att-kgcn: Tourist attractions recommendation system by using attention mechanism and knowledge graph convolution network. In *2023 3rd International Conference on Emerging Smart Technologies and Applications (eSmarTA)*. IEEE.

Curtis Murray, Lewis Mitchell, Jonathan Tuke, and Mark Mackay. 2024. Probabilistic emotion and sentiment modelling of patient-reported experiences. *Preprint*, arXiv:2401.04367.

Kabir Nagrecha, Lingyi Liu, Pablo Delgado, and Prasanna Padmanabhan. 2023. Intune: Reinforcement learning-based data pipeline optimization for deep recommendation models. *Preprint*, arXiv:2308.08500.

Oron Nir, Idan Vidra, Avi Neeman, Barak Kinarti, and Ariel Shamir. 2024. Vcr: Video representation for contextual retrieval. *Preprint*, arXiv:2402.07466.

Stephen Obadinma, Alia Lachana, Maia Norman, Jocelyn Rankin, Joanna Yu, Xiaodan Zhu, Darren Mastropaolo, Deval Pandya, Roxana Sultan, and Elham Dolatabadi. 2024. Faiir: Building toward a conversational ai agent assistant for youth mental health service provision. *Preprint*, arXiv:2405.18553.

Ayano Okoso, Keisuke Otaki, Satoshi Koide, and Yukino Baba. 2024. Impact of tone-aware explanations in recommender systems. *Preprint*, arXiv:2405.05061.

Rickson Simioni Pereira, Claudio Di Sipio, Martina De Sanctis, and Ludovico Iovino. 2024. On the need for configurable travel recommender systems: A systematic mapping study. *Preprint*, arXiv:2407.11575.

Dimitrios Rafailidis. 2019. Leveraging trust and distrust in recommender systems via deep learning. *Preprint*, arXiv:1905.13612.

Jiarui Rao and Jionghao Lin. 2024. Ramo: Retrieval-augmented generation for enhancing moocs recommendations. *Preprint*, arXiv:2407.04925.

Aadirupa Saha and Pierre Gaillard. 2024. Stop relying on no-choice and do not repeat the moves: Optimal, efficient and practical algorithms for assortment optimization. *Preprint*, arXiv:2402.18917.

Srijan Saket, Olivier Jeunen, and Md. Danish Kalim. 2024. Monitoring the evolution of behavioural embeddings in social media recommendation. *Preprint*, arXiv:2312.15265.

Navid Seidi. 2024. A stable matching assignment for cancer treatment centers using survival analysis. *Preprint*, arXiv:2401.10469.

Dinesh Cyril Selvaraj, Falko Dressler, and Carla Fabiana Chiasserini. 2024. Personalized and context-aware route planning for edge-assisted vehicles. *Preprint*, arXiv:2407.17980.

Lütfi Kerem Senel, Besnik Fetahu, Davis Yoshida, Zhiyu Chen, Giuseppe Castellucci, Nikhita Vedula, Jason Choi, and Shervin Malmasi. 2024. Generative explore-exploit: Training-free optimization of generative recommender systems using llm optimizers. *Preprint*, arXiv:2406.05255.

Apurva Sinha and Ekta Gujral. 2024. Pae: Llm-based product attribute extraction for e-commerce fashion trends. *Preprint*, arXiv:2405.17533.

Yannik Stein and Fabian Moerchen. 2020. A general approach to solving bandit problems. Amazon Science.

Tom Sühr, Samira Samadi, and Chiara Farronato. 2024. A dynamic model of performative human-ml collaboration: Theory and empirical evidence. *Preprint*, arXiv:2405.13753.

Chun How Tan, Austin Chan, Malay Haldar, Jie Tang, Xin Liu, Mustafa Abdool, Huiji Gao, Liwei He, and Sanjeev Katariya. 2023. Optimizing airbnb search journey with multi-task learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23. ACM.

Samuel Tan and Peter I. Frazier. 2024. Asymptotically optimal regret for black-box predict-then-optimize. *Preprint*, arXiv:2406.07866.

Vishnu Unnikrishnan, Clara Puga, Miro Schleicher, Uli Niemann, Berthod Langguth, Stefan Schoisswohl, Birgit Mazurek, Rilana Cima, Jose Antonio Lopez-Escamez, Dimitris Kikidis, Eleftheria Vellidou, Ruediger Pryss, Winfried Schlee, and Myra Spiliopoulou. 2024. Training and validating a treatment recommender with partial verification evidence. *Preprint*, arXiv:2406.06654.

Ghanshyam Verma, Shovon Sengupta, Simon Simanta, Huan Chen, Janos A. Perge, Devishree Pillai, John P. McCrae, and Paul Buitelaar. 2023. Empowering recommender systems using automatically generated knowledge graphs and reinforcement learning. *Preprint*, arXiv:2307.04996.

Alicia Vidler. 2024. Recommender systems in financial trading: Using machine-based conviction analysis in an explainable ai investment framework. *Preprint*, arXiv:2404.11080.

Karan Vombatkere, Sepehr Mousavi, Savvas Zannettou, Franziska Roesner, and Krishna P. Gummadi. 2024. Tiktok and the art of personalization: Investigating exploration and exploitation on social media feeds. *Preprint*, arXiv:2403.12410.

Junting Wang, Adit Krishnan, Hari Sundaram, and Yunzhe Li. 2023. Pre-trained neural recommenders: A transferable zero-shot framework for recommendation systems. *Preprint*, arXiv:2309.01188.

Ruofan Wang, Prakruthi Prabhakar, Gaurav Srivastava, Tianqi Wang, Zeinab S. Jalali, Varun Bharill, Yunbo Ouyang, Aastha Nigam, Divya Venugopalan, Aman Gupta, Fedor Borisyuk, Sathiya Keerthi, and Ajith Muralidharan. 2024. Limaml: Personalization of deep recommender models via meta learning. *Preprint*, arXiv:2403.00803.

Shen Wang, Jibing Gong, Jinlong Wang, Wenzheng Feng, Hao Peng, Jie Tang, and Philip S. Yu. 2020. Attentional graph convolutional networks for knowledge concept recommendation in moocs in a heterogeneous view. *Preprint*, arXiv:2006.13257.

Yanfang Wang, Yongduo Sui, Xiang Wang, Zhenguang Liu, and Xiangnan He. 2022a. Exploring lottery ticket hypothesis in media recommender systems. *International Journal of Intelligent Systems*, 37(5):3006–3024.

Yongwei Wang, Yong Liu, and Zhiqi Shen. 2022b. Revisiting item promotion in gnn-based collaborative filtering: A masked targeted topological attack perspective. *Preprint*, arXiv:2208.09979.

Hao Wu, Alejandro Ariza-Casabona, Bartłomiej Twardowski, and Tri Kurniawan Wijaya. 2023. Mm-gef: Multi-modal representation meet collaborative filtering. *Preprint*, arXiv:2308.07222.

Da Xu, Danqing Zhang, Guangyu Yang, Bo Yang, Shuyuan Xu, Lingling Zheng, and Cindy Liang. 2024a. Survey for landing generative ai in social and e-commerce recsys – the industry perspectives. *Preprint*, arXiv:2406.06475.

Xiaonan Xu, Yichao Wu, Penghao Liang, Yuhang He, and Han Wang. 2024b. Emerging synergies between large language models and machine learning in ecommerce recommendations. *Preprint*, arXiv:2403.02760.

Xiaonan Xu, Yichao Wu, Penghao Liang, Yuhang He, and Han Wang. 2024c. Emerging synergies between large language models and machine learning in ecommerce recommendations. *Preprint*, arXiv:2403.02760.

Hongrui Xuan, Yi Liu, Bohan Li, and Hongzhi Yin. 2023. Knowledge enhancement for contrastive multi-behavior recommendation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23. ACM.

Michiharu Yamashita, Thanh Tran, and Dongwon Lee. 2024. Fake resume attacks: Data poisoning on online job platforms. *Preprint*, arXiv:2402.14124.

Xiaodong Yang, Huiyuan Chen, Yuchen Yan, Yuxin Tang, Yuying Zhao, Eric Xu, Yiwei Cai, and Hanghang Tong. 2024. Simce: Simplifying cross-entropy loss for collaborative filtering. *Preprint*, arXiv:2406.16170.

Kelley Ann Yohe. 2023. Towards global, socio-economic, and culturally aware recommender systems. *Preprint*, arXiv:2312.05805.

Savvas Zannettou, Olivia-Nemes Nemeth, Oshrat Ayalon, Angelica Goetzen, Krishna P. Gummadi, Elissa M. Redmiles, and Franziska Roesner. 2024. Analyzing user engagement with tiktok's short format video recommendations using data donations. *Preprint*, arXiv:2301.04945.

Shun Zhang, Runsen Zhang, and Zhirong Yang. 2024a. Matrrec: Uniting mamba and transformer for sequential recommendation. *Preprint*, arXiv:2407.19239.

Yi Zhang, Lei Sang, and Yiwen Zhang. 2024b. Exploring the individuality and collectivity of intents behind interactions for graph collaborative filtering. *Preprint*, arXiv:2405.09042.

Zihao Zhao, Yi Jing, Fuli Feng, Jiancan Wu, Chongming Gao, and Xiangnan He. 2024. Leave no patient behind: Enhancing medication recommendation for rare disease patients. *Preprint*, arXiv:2403.17745.

Zhi Zheng, Wenshuo Chao, Zhaopeng Qiu, Hengshu Zhu, and Hui Xiong. 2024. Harnessing large language models for text-rich sequential recommendation. *Preprint*, arXiv:2403.13325.

Jieming Zhu, Chuhan Wu, Rui Zhang, and Zhenhua Dong. 2024a. Multimodal pretraining and generation for recommendation: A tutorial. *Preprint*, arXiv:2405.06927.

Yongchun Zhu, Jingwu Chen, Ling Chen, Yitan Li, Feng Zhang, and Zuotao Liu. 2024b. Interest clock: Time perception in real-time streaming recommendation system. *ArXiv*, abs/2404.19357.

Yongchun Zhu, Yudan Liu, Ruobing Xie, Fuzhen Zhuang, Xiaobo Hao, Kaikai Ge, Xu Zhang, Leyu Lin, and Juan Cao. 2021. Learning to expand audience via meta hybrid experts and critics for recommendation and advertising. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21. ACM.

Siyu Zhuo, Xiaoning Zhu, Pan Shang, and Zhengke Liu. 2024. Passenger route and departure time guidance under disruptions in oversaturated urban rail transit networks. *Preprint*, arXiv:2407.03388.