User Intent Modeling

VIPASHA VAGHELA, DAIICT, India
ARPIT RANA, DAIICT, India

A clear and well-documented LaTeX document is presented as an article formatted for publication by ACM in a conference proceedings or journal publication. Based on the "acmart" document class, this article presents and explains many of the common variations, as well as many of the formatting elements an author may use in the preparation of the documentation of their work.

Additional Key Words and Phrases: Recommendation

ACM Reference Format:
Vipasha Vaghela and Arpit Rana. 2023. User Intent Modeling. 1, 1 (October 2023), 21 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

User intent modeling is a critical component of recommendation systems, particularly in the realm of personalized content recommendations. It plays a pivotal role in understanding users' underlying goals and objectives when interacting with a system or performing searches. This understanding enables recommendation systems to provide tailored content or product suggestions that match the user's intent, enhancing personalization and the overall customer journey. User intent can take various forms, including informational, commercial, navigational, or transactional, and accurately discerning these intents is crucial for delivering relevant recommendations. Even in cases of ambiguous or evolving intent, session-based personalization helps adapt recommendations to the user's current context, considering factors like location and device type. In essence, user intent modeling empowers recommendation systems to deliver more accurate, context-aware, and user-centric suggestions, contributing to a superior user experience and effective content strategies.

2 USER INTENT MODELING TAXONOMY

User intent modeling is a crucial aspect of recommendation systems, aiming to understand and predict what users want or intend to do. Various methods and techniques are employed to model user intent, we basically devide this into 3 methods:

2.1 Session Agnostic Modeling

Session agnostic user intent modeling focuses on capturing and analyzing users' historical behavior and browsing history without considering session-specific information. It aims to understand users' preferences and intent over an extended period, providing a holistic view of their long-term interactions with the system.

Aspect-based intent representation, on the other hand, is an approach used in recommendation systems to comprehensively capture and represent the various dimensions or aspects of a user's

Authors' addresses: Vipasha Vaghela, DAIICT, Ahmedabad, Gujarat, India; Arpit Rana, DAIICT, Ahmedabad, Gujarat, India.

XXXX-XXXX/2023/10-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

preferences that have developed over time. It involves categorizing and structuring a user's evolving preferences across different aspects or categories of items over an extended period. By doing so, it enables recommendation systems to provide more accurate and personalized recommendations that align closely with the user's preferences and motivations, resulting in a more satisfying user experience.

2.1.1 ***Aspect Based User Profile Representation***.

User profile modeling creates computational models to predict user needs and preferences. User profiles contain information about rules, settings, interests, and more, which can be static or dynamic. Accuracy depends on data collection and organization. In recommendation systems, user profiles enhance personalized experiences.

Two main user modeling approaches exist: initial profiles and ongoing updates. Ensuring accurate profiles is challenging, particularly for new users ("cold start"). Dynamic user models that adapt to changing interests are crucial.

Aspect-based profile modeling divides user preferences into specific dimensions. It tailors recommendations by understanding the user's interests in each dimension. This is valuable in areas like e-commerce and content recommendation for precise, relevant suggestions.

Item Based Profile Representation

n recommendation systems, item-based intent representation is a valuable technique used to describe items by examining their intrinsic qualities or attributes, aiming to enhance the precision and personalization of recommendations. Unlike user-centric intent representation, which focuses on understanding user preferences, item-based intent representation concentrates on defining the items themselves. This approach involves creating detailed profiles of items to improve the accuracy and relevance of recommendations.

In recommendation systems, item-based intent representation is a valuable technique used to describe items by examining their intrinsic qualities or attributes, aiming to enhance the precision and personalization of recommendations. Unlike user-centric intent representation, which focuses on understanding user preferences, item-based intent representation concentrates on defining the items themselves. This approach involves creating detailed profiles of items to improve the accuracy and relevance of recommendations. Aspect-based intent representation is a fundamental concept within recommendation systems, and it plays a pivotal role in enhancing the accuracy and relevance of recommendations over an extended period. This approach involves capturing and representing various aspects or dimensions of a user's preferences, enabling a comprehensive understanding of their long-term preferences. By structuring the comprehension of a user's evolving preferences and capturing their intents across different aspects or categories of items, this representation aligns recommendations closely with the user's preferences and motivations.

In the context of recommendation systems, accurately understanding user intent is a challenge, particularly when users do not explicitly convey their preferences. [3] introduces the concept of a "user aspect," akin to a query intent in recommender systems. In such systems, users rely on the system to infer their preferences based on past interactions, a situation resembling the ambiguity in information retrieval. Determining the user's current intent for a given interaction can be challenging, and making an incorrect assumption about user intent can lead to recommendations that users do not find relevant.

To address these challenges, [2] presents a Probabilistic Latent Semantic Analysis model that not only utilizes explicit aspects but also learns aspect probabilities directly. This approach

strengthens the intent-aware framework by ensuring that the aspect probabilities, crucial for the method, accurately reflect real user preferences. While maintaining the interpretability of the co-occurrence counting method, which aids in understanding known aspects, latent aspects, while effective for predictive performance, lack interpretability. [2] proposes a compromise by employing an aspect model based on explicit aspects and enhancing it through learning the model components to optimize predictive performance, thereby achieving a balance between interpretability and prediction accuracy.

Users often exhibit a diverse set of interests, spanning various types of items they enjoy. For instance, a user primarily interested in cartoon movies may occasionally explore family and action films. Effective recommendations require providing a diversified list of suggested items that cater to these varied interests. In this context, recommending a mix of cartoons, action, and family movies, rather than solely focusing on cartoon movies, is preferable. Additionally, users sometimes engage in exploratory journeys across different interests without a specific goal in mind.

In summary, aspect-based intent representation is a critical element in recommendation systems, enabling precise and personalized recommendations by capturing long-term user preferences. Addressing challenges related to understanding user intent and achieving a balance between interpretability and prediction accuracy are key considerations in the field. Users engage with items for various intents, which can be latent or hidden, making their effective use in Sequential Recommendation challenging.

Many recommendation models primarily rely on a user's item-level interaction history to discern their topical interests. However, these models often lack a deeper understanding of user intent, such as the user's specific goals or desires at a given moment—whether it be discovering new content, resuming from a previous session, or enjoying background music. Understanding user intents is pivotal for enhancing long-term user experiences.

In summary, these advancements in intent representation and recommendation algorithms contribute to more accurate, personalized, and diverse recommendations, addressing the challenges associated with understanding user intent, capturing latent preferences, and enhancing recommendation transparency and explainability.

Rating Based Profile Representation

In a study presented in [7], a personalized recommendation algorithm is introduced to address issues of low recommendation accuracy and the "cold start" problem in traditional recommendation systems. This algorithm is based on user profiles and user preferences and outperforms traditional collaborative filtering methods. It begins by establishing a user model using historical rating data and incorporates user profile information in the calculation of user similarity. Further, it adjusts user profile parameters to optimize the user model based on individual profile differences among users. Experimental results demonstrate the effectiveness of this personalized recommendation algorithm in improving recommendation system performance.

Additionally, many existing recommendation models fail to consider spatial information related to users and items. To mitigate the cold start problem and enhance recommendation accuracy, a location-aware probabilistic generative model is proposed in [8]. This model leverages location-based ratings to create user profiles and generate recommendations. By incorporating user and item locations associated with ratings, it offers a promising way to provide context-aware recommendations. This approach contributes to a more comprehensive understanding of user preferences by factoring in real-world location context, thus addressing the limitations of traditional recommendation models in handling user ratings.

Category Based Profile Representation

A category-based user profile is a representation of a user's preferences or interests based on the categories or types of items they have interacted with or shown a preference for in the past. For example, in a movie recommendation system, a category-based user profile might represent a user's preferences for different movie genres, such as action, comedy, or romance. If a user has rated or interacted with several action movies in the past, the category-based user profile would indicate a strong preference for the "action" category.

In the context of category-based user profile modeling, the objective is to create a more robust and context-rich personalization experience for users. This is achieved by developing a user interest hierarchy that encompasses a spectrum of interests, ranging from broad to specific, without requiring direct user input or feedback, Learn a hierarchy of topics of interests from history of user to provide a context for personalization with topic clustering [14].The approach taken in this endeavor is similar to clustering, where the hierarchical category structure is formed as a keyword vector representation of a user's interests. This hierarchy is designed to provide a comprehensive context for personalization, implicitly learning from the user's interaction with content, particularly bookmarked web pages [14].

Moreover, the problem of over-personalization in recommender systems is addressed by introducing a novel category-based user model [13]. This model is aimed at diversifying user profiles, focusing on categorical preferences within the same user group during recommendation filtering. Over-personalization arises when recommendations become excessively tailored to individual preferences, potentially reducing user satisfaction. It is a consequence of prioritizing accuracy metrics at the expense of user experience [13].

To combat over-personalization, the proposed solution emphasizes diversity in recommendations across various dimensions, including global coverage, local coverage, novelty, and redundancy. By diversifying user profiles within the same group, considering categorical preferences, this approach strives to enhance the recommendation experience. It introduces a user-centric diversification framework and a novel category-based user model that categorizes users based on their preferences for specific item categories [13].

Furthermore, the paper introduces a method for modeling user preferences based on categories, involving the extraction of positive ratings from individual users. This modeling approach takes into account average ratings and individual thresholds to eliminate biases associated with rating scales. While primarily designed for explicit ratings, the paper suggests its potential applicability to implicit ratings as well [13].

2.1.2 User Sub Profile Representation.

In many approaches to recommendation diversification, a recommender scores items for relevance and then re-ranks them to balance relevance with diversity,presented a personalized form of intent-aware diversification, in which the aspects to be covered by the reranked recommendations are subprofiles of the user's profile, each representing a distinct user taste. [4],[5],[11] addresses the challenge of balancing relevance and diversity in recommendation systems. It presents a personalized intent-aware diversification approach called SubProfile-Aware Diversification, uses user subprofiles, which represent distinct user tastes and are derived from a user's profile. Unlike conventional methods that rely on item features, adapts to situations where item features are missing or of low quality.It improves recommendation personalization and diversity by re-ranking top recommendations and identifying user subprofiles based on these suggestions. It operates on positively-rated items in a user's profile and eliminates redundant explanations to determine the subprofiles. This method outperforms traditional approaches that depend on item features, offering a better balance between recommendation accuracy and diversity.

2.2 Session Based Modeling

Session-Based Intent Modeling is a recommendation system approach that focuses on understanding a user's immediate intent and preferences during their current session, as opposed to considering long-term historical interactions. It analyzes the real-time behaviors and actions within a single user session, such as clicks and searches, to provide immediate and relevant recommendations that align with the user's current session intent. This approach is valuable in scenarios where user preferences change rapidly or where users have varied interests within a single session, and it finds applications in e-commerce, content streaming, and online news domains. It enhances the user experience by ensuring that recommendations are timely, context-aware, and aligned with the user's immediate interests.

2.2.1 Intra Session Modeling.

Intra-session user intent modeling is the process of understanding and predicting a user's intentions and preferences within a single interaction with a digital system, focusing on short-term behaviors and decisions during a session rather than long-term preferences. This approach is particularly valuable in domains like e-commerce and content recommendation, where timely delivery of relevant information is essential for user engagement and satisfaction.

To address the challenges of short web user sessions and data availability for short-term user modeling, a novel approach is proposed in [23]. This approach captures changes in user behavior by comparing current session data with previous sessions, making it applicable to existing systems with readily available data sources.

In session-based modeling, user intent and preference play crucial roles in influencing a user's decisions. Existing approaches often oversimplify user intent by assuming a single dominant intent within a session. To overcome this limitation, a comprehensive framework is introduced in [43], explicitly modeling user intent and preference. This framework, featuring a key-array memory network with a hierarchical intent tree, empowers to differentiate between user intent and preference, addressing issues related to multiple intents within a session.this model captures intent-driven user behavior more effectively and provides intent-specific user embeddings.

In summary, these advancements in intra-session user intent modeling, as discussed in [23] and [43], address the challenges of short web user sessions and provide more nuanced and personalized recommendations. These approaches enhance the overall user experience by adapting to users' evolving needs and preferences within the constraints of brief interactions, ultimately leading to more effective and satisfying digital experiences.

Time Aware Intent Modeling

Traditional recommendation approaches often overlook the temporal context of user preferences, posing challenges in accurately predicting short-term interests. To address this limitation, a time-aware distributional content-based recommender system is introduced in [17]. This system focuses on suggesting items that users will likely be interested in shortly, leveraging an exponential decay function and semantic similarity between item descriptions to regulate item participation in building user profiles. Items are represented in a WordSpace, where related words are represented as near points (vectors), and the semantics of item descriptions are computed by summing the vectors associated with their words.

While Session-Based Recommender Systems (SBR) like the one described in [21] predict users' next items based on their recent session interactions, they often lack user identity information and primarily cater to short-term preferences. Unlike other recommender systems, they cannot employ user-item matrix-analysis models and place significant emphasis on current session interactions. However, one underexplored aspect within SBR data is the role of time intervals between user-item interactions. These intervals carry meaningful signals,

potentially signifying factors like item importance or user distraction. For instance, longer time intervals might indicate higher interest.

To harness the potential of these time intervals, the STAR framework is introduced, as mentioned in [21]. STAR enhances SBRs by incorporating time intervals between events within sessions without discretization. It constructs session representations that reflect users' current interests and employs them to predict their next preferences. This framework leverages temporal information to enhance recommendation quality, identifying discontinuities in session events and adjusting item weights based on time intervals. Experimental results demonstrate that considering temporal information efficiently captures user interests during sessions, opening up avenues for further research.

Furthermore, [22] introduces the Time-Aware Neural Attention Network for Session-Based Recommender Systems (TNARM) to address issues related to inaccurate user preference models and subpar recommendations in existing systems due to the neglect of click times. TNARM constructs a global session graph, utilizes Graph Neural Networks (GNN) to learn item embeddings, refines these embeddings using a Gated Recurrent Unit (GRU) to capture user interests in the current session, and introduces a time-aware neural attention network to identify the user's main session purpose. These components synergize to create dynamic user preferences, ultimately leading to improved recommendations.

In summary, these innovations in recommender systems, as highlighted in [17], [21], and [22], collectively address the challenge of capturing and utilizing temporal information in user preference modeling. These approaches enhance the accuracy of short-term interest predictions and contribute to more effective recommendations, ultimately enriching the user experience in various application domains.

Transitional relation Aware Modeling

Transition Relation Aware Modeling refers to a technique used in recommendation systems to understand and model the intricate and complex session transition relationships within a user's interactions. This approach focuses on recognizing how users move from one state or context to another during their interactions with a system or platform. By capturing these complex session transition patterns, recommendation systems can provide more accurate and context-aware suggestions, taking into account the evolving preferences and interests of users as they navigate through various stages or contexts within a session. Transition Relation Aware Modeling is particularly valuable for improving the quality of recommendations in sequential recommendation tasks where the sequence and transitions of user actions play a crucial role in understanding user intent.

Graph Neural Networks are essential for Transition Relation Aware Modeling in recommendation systems, allowing the effective capture and utilization of complex session transition relationships. This modeling approach aims to understand how users navigate through various states or contexts during their interactions. GNNs excel in modeling intricate transitions within sessions, enabling recommendation systems to provide more accurate and context-aware recommendations as users progress through sessions. [27] In this passage, the problem of session-based recommendation is addressed, where the challenge is predicting user actions within anonymous sessions, For instance, Session-Based Recommendation with [27] Graph Neural Networks (SR-GNN) utilizes GNNs to build session graphs based on item transition relations, leveraging a Gated Graph Neural Network to extract these transitions. extended this approach by incorporating a GGNN for local context and a self-attention network for global dependencies. Despite their promise, GNN-based methods, such as these, have faced challenges in effectively encoding session information.stands out by transforming session

sequences into graph structures and employing Graph Neural Networks to capture complex item transitions. SR-GNN represents each session using a blend of global preferences and session-specific interests, avoiding the need for accurate user representations. Extensive experiments indicate that SR-GNN outperforms existing methods, making it a promising approach for session-based recommendation by enhancing accuracy and addressing key limitations.

In recent studies, several innovative approaches have been introduced to enhance session-based recommendation systems, each addressing specific challenges in this domain. First, a novel approach introduces multi-granularity consecutive user intent units, employing the Multi-granularity Intent Heterogeneous Session Graph (MIHSG) to capture interactions between different granularity intent units. This method aims to alleviate issues related to long-range dependencies in recommendation. Furthermore, an Intent Fusion Ranking (IFR) module is proposed to combine recommendation results from various granularity user intents. The MIHSG leverages a heterogeneous graph to model transition relationships among these intent units, while the IFR strategy leverages intent representations from all granularity levels, as discussed in [24].

Another challenge in session-based recommendation is understanding session context, which is crucial for accurate user preference prediction. To address this, a two-step approach is introduced. It generates session embeddings from a user-item multigraph and clusters sessions to identify session context clusters. This technique utilizes a session-item bipartite multigraph and employs the GraphSage node embedding method to create session embeddings, as highlighted in [25].

Additionally, some session-based recommender systems now consider user price preferences alongside interest preferences. The Co-guided Heterogeneous Hypergraph Network (CoHHN) addresses this challenge by modeling both user interest and price preferences in session-based recommendation. CoHHN uses a heterogeneous hypergraph to represent various information, including item price, ID, and category, linking them to price and interest preferences. This approach involves three types of hyperedges and employs a dual-channel aggregating mechanism and attention layers to extract preferences effectively. CoHHN introduces a co-guided learning scheme to enhance the interplay between price and interest preferences, ultimately leading to improved recommendations, as described in [26].

Finally, session-based recommendation systems, known for their practicality and privacy-friendliness, focus on predicting users' next actions within sessions. To address the limitations of early methods, the CaSe4SR (Category Sequence graph for Session-based Recommendation) approach is introduced. It leverages item category information to enhance session-based recommendation. It constructs item and category graphs from user behavior and category sequences to reduce noise and provide clearer user interest signals. This model employs Graph Neural Networks to learn representations, fusion strategies, and an attention mechanism, as explained in [27].

In summary, these pioneering approaches collectively enhance session-based recommendation systems by tackling various challenges, including session encoding, session context modeling, the consideration of user price preferences, and the improved utilization of auxiliary information like item categories, resulting in more accurate and effective recommendations.

2.2.2 Inter Session Modeling.

Inter-session modeling is an approach used in recommendation systems to understand and predict user preferences, behavior, and intent across multiple sessions or interactions with a digital platform.

In this modeling approach, the system looks at the broader picture of a user's interactions and preferences over time, beyond just a single session or session boundary.

Inter-session modeling in the context of session-based recommender systems (SBRS) is a vital approach aimed at understanding user preferences and intent over multiple sessions or interactions with a digital platform, such as a website or application. While traditional session-based modeling primarily focuses on predicting a user's next action within a single session, inter-session modeling extends this understanding to encompass the user's behavior and preferences across various sessions, often spanning a more extended period. It plays a crucial role in improving the accuracy of recommendations and personalization by considering a user's long-term behavior and preferences.

Linear Sequential Modeling

Traditional Recurrent Neural Networks (RNNs) have shown promise in capturing sequential patterns in session-based settings, they often focus on modeling user behavior within a single session, also inter-session information. Recurrent Neural Networks (RNNs), as demonstrated in [19]. RNNs are effective in capturing sequential patterns and have shown promise in session-based settings. This approach models entire sessions rather than relying on item-to-item recommendations, which can lead to more accurate recommendations. Modifications to classic RNNs, such as introducing a ranking loss function, are employed to adapt them for this specific problem. However, this approach often focuses on modeling session data within a single session and may not consider inter-session information, limiting its ability to capture long-term user intent.There is another approach using Recurrent Neural Network. In a hierarchical RNN model proposed by [18], the hidden states of RNNs evolve across user sessions. Specifically, the hidden state of a lower-level RNN at the end of one user session is passed as input to a higher-level RNN. This higher-level RNN then predicts an initialization (context vector) for the hidden state of the lower RNN for the next user session. This approach enables the modeling of inter-session information transfer, allowing the system to capture the evolution of user intent over time. Another innovative approach, as presented by [20], introduces the concept of an Inter-Intra RNN (II-RNN) for session-based recommendations. In this model, both inter-session and intra-session behavior modeling are combined within a single architecture. The inter-session RNN provides the initial hidden state for the intra-session RNN, enhancing the system's ability to make informed recommendations even at the start of a session. This approach aims to address the "cold start" problem, where traditional recommendation models struggle to provide accurate recommendations early in a session due to limited data. The II-RNN model leverages the temporal order of user actions, making it well-suited for session-based recommendations. In conclusion, inter-session modeling is a critical component of recommendation systems that aims to capture the evolution of user intent and preferences over time. Techniques such as hierarchical RNNs, Inter-Intra RNNs, and GNN-based frameworks offer effective ways to model and leverage inter-session information, ultimately leading to more accurate and personalized recommendations across multiple user interactions and sessions.

Non Sequential Dependency Modeling

[28]non-sequential modeling specifically focusing on the use of Graph Neural Networks as an alternative approach. GNNs are known for their ability to model complex user behavior transitions and item dependencies in session-based settings. In this approach, each item within a session is treated as a node, and consecutive nodes are connected by edges, effectively transforming the session data into a graph structure. [30] shows the popularity of GNNs in session-based recommendation systems, [30] particularly This recommendation framework employs three GNNs to capture both intra-session and inter-session item correlations, as

well as session-session correlations. By considering these different levels of correlations, it generates session representations and aims to address data sparsity challenges by expanding node items in a global item graph constructed from all training sessions.
However, [28] highlights a limitation of standard GNN-based models. These models typically consider all neighboring nodes as equally influential, which may not accurately represent the nuanced relationships between items. In a real-world scenario, items within a session may have varying levels of relevance or importance to a user, and treating them equally may not be optimal. To address this limitation, [28] introduces a concept called a "multi-interest graph." This approach takes into account item dependencies in relation to various user interests. By doing so, it allows for more precise and interest-based recommendations. In other words, it recognizes that not all items are of equal importance to a user and aims to make recommendations that are tailored to the specific interests and preferences of the individual user, rather than relying solely on the sequential order of items within a session.
[29] introduces innovative approaches, which modify the traditional session-based recommendation methods to better capture complex non-sequential item dependencies and enhance the modeling of inter-session information, using Full Graph Neural Networks to learn intricate item dependencies. Rather than representing sessions as linear sequences, each session is transformed into a graph structure, which can better capture complex non-sequential item dependencies. To incorporate inter-session information, a Broadly Connected Session graph links different sessions, and a Mask-Readout function enhances session embeddings based on the graph. This approach addresses the challenge of session anonymity and allows for the consideration of data from multiple sessions, further enhancing recommendation accuracy.

2.3 Session Aware Modeling

Session-aware modeling represents a sophisticated approach to recommendation systems that take into account both short-term session context and long-term user preferences. These systems recognize that user interests and preferences can evolve over time, and they aim to provide more relevant recommendations by considering both the user's current interactions within a session and their historical behavior across multiple sessions.

2.3.1 Intent Modeling in Conventional Single Shot Recommender.

A single-shot conventional recommendation system is a type of recommendation system that provides users with personalized recommendations in a single interaction or query without the need for ongoing user interactions or feedback. Unlike some recommendation systems that rely on continuous user feedback to improve recommendations over time, a single-shot recommendation system aims to make accurate recommendations immediately based on the available data and user profile. It typically uses various algorithms and techniques, such as collaborative filtering, content-based filtering, or hybrid approaches, to generate recommendations without the need for an extended history of user actions or preferences. This type of system is suitable for scenarios where users may not have a long history of interactions with the platform or where real-time recommendations are essential.

Pattern Based Modeling

Pattern-Based Intent Modeling is an approach utilized in recommendation systems and user modeling to discern and anticipate user intent by analyzing recurring behavioral patterns. This method concentrates on identifying repetitive sequences or trends in user actions, such

as clicks, searches, or interactions, which offer valuable insights into user preferences and intentions. Let's delve into the intricacies of Pattern-Based Intent Modeling with reference to the following research papers:

In a bid to enhance the performance of recommender systems, [31] addresses prevalent challenges like data sparsity, long-tailed datasets, and cold-start issues. The approach advocated involves the amalgamation of information from related events, such as social connections or sequences of recent activities. This supplementary information is harnessed to augment the efficacy of recommendation techniques based on matrix factorization. [31] introduces novel methods that concurrently leverage social and sequential data, with the aim of bolstering recommendation performance, particularly in scenarios characterized by data sparsity and cold-start challenges. The former posit that user actions are influenced by their recent activities, while the latter posit that user actions can be predicted based on the behavior of their friends. Both approaches employ related activities as a form of regularization to refine predictions when data is insufficient to model users and items independently. [31] distinguishes between two existing models, Factorized Personalized Markov Chains and Social Bayesian Personalized Ranking, which cater to these two approaches. FPMC enhances overall performance but falls short in addressing cold-start problems due to the substantial parameter requirements. In contrast, SBPR prioritizes items based on user and friend interactions, thus achieving increased accuracy for cold-start scenarios.

To resolve these challenges, the paper introduces a novel model named SPMC Socially-aware Personalized Markov Chain. SPMC marries feedback from sequences and social interactions within a unified framework, underpinned by the assumption that user behavior is influenced by their preferences, recent actions, and the recent actions of their friends. This innovative approach aspires to surpass existing models, particularly in mitigating cold-start issues within sparse datasets. SPMC distinguishes itself by its sequential awareness, which enables successive predictions—a desirable trait for recommender systems. Moreover, it factors in the influence of users' friends' recent activities on their future actions, steering clear of the presumption of long-term preferences within social circles. The paper substantiates the effectiveness of modeling socio-temporal dynamics through empirical comparisons with state-of-the-art socially-aware recommendation methods.

Turning to [32], it explores the pivotal role of recommender systems on various websites, spotlighting two primary approaches: matrix factorization (MF) and Markov chains (MC). MF endeavors to ascertain a user's general preferences by factoring the matrix of observed user-item interactions. Conversely, MC methods focus on modeling sequential behavior, deciphering a transition matrix over items to predict a user's next action predicated on their recent activities. The paper introduces a pioneering method known as Factorized Personalized MC (FPMC), which effectively marries these two approaches. FPMC pioneers personalized MCs, thereby bridging the gap between MC and MF. Instead of relying on a uniform transition matrix across all users, FPMC crafts a personalized transition cube, with each slice representing a user-specific transition matrix derived from the user's basket history. This personalized approach encapsulates both sequential effects and long-term user preferences. The paper introduces a factorisation model for the transition tensor to address data sparsity and ensure dependable estimates of personalized transition matrices. This model propagates information among similar users, items, and transitions, effectively mitigating sparsity issues. Furthermore, the paper extends the Bayesian Personalized Ranking (BPR) framework to cater to basket data, ensuring robust parameter learning. The efficacy of FPMC is substantiated through real-world e-commerce dataset evaluations, demonstrating its superiority over traditional MF and MC models. Key contributions encompass the introduction

of personalized Markov chains, the development of a factorization model to combat sparsity, and the extension of the BPR framework to accommodate basket data.
Lastly, [33] delves into recommender systems, illuminating the coexistence of two primary paradigms: Matrix Factorization (MF) and Markov Chains (MC). MF methods aim to decipher a user's long-term preferences through matrix factorization, whereas MC methods center around modeling sequential behavior and learning a transition graph over items. The paper introduces FactOrized Sequential Prediction with Item SImilarity ModeLs (Fossil) as an innovative approach that adeptly melds these paradigms, especially in scenarios featuring sparse datasets. Fossil effectively combines similarity-based methods such as Factored Item Similarity Models (FISM) with MC methods akin to FPMC, thus accommodating sparse datasets replete with sequential dynamics. It does so by introducing a personalized weighting scheme over item sequences, effectively characterizing users in terms of their preferences and sequential behavior. Fossil offers several advantages for mitigating sparsity issues, such as parameterizing users based on historical items to mitigate cold-user issues and adapting to users with limited historical data by accentuating short-term dynamics, leveraging global sequential patterns.
In summation, Pattern-Based Intent Modeling, represents a dynamic approach to understanding and anticipating user intent by recognizing recurrent behavioral patterns. These papers underscore the significance of combining diverse data sources, tackling cold-start challenges, and addressing sparsity issues to enhance the precision and personalization of recommender systems.

Model Based Modeling

Model-based user intent modeling is an approach used in recommendation systems to understand and predict user intent, preferences, and behaviors through the use of computational models. This method involves creating mathematical or algorithmic models that analyze and represent user data to make predictions about a user's future actions or interests. These models are typically based on patterns and relationships extracted from historical user interactions and data.

- **Single Intent Modeling**
  User single intent modeling is a concept in recommendation systems that focuses on understanding and predicting a user's primary or dominant intention or interest within a specific context. In this modeling approach, the system aims to identify and prioritize a single intent or preference that is most relevant to the user's current session or interaction. This is in contrast to multi-intent modeling, which considers multiple user intentions simultaneously.
  User single intent modeling is valuable in scenarios where users have a clear and primary objective, such as making a purchase, finding specific information, or watching a particular type of content. By focusing on the user's primary intent, recommendation systems can provide more accurate and relevant suggestions, enhancing the user's experience and achieving the user's primary goal.
  This modeling approach can be applied in various domains, including e-commerce, content recommendation, and search engines, where understanding and catering to the user's primary intent is crucial for delivering satisfying and efficient recommendations.

  – **Unidirectional Modeling**
    Unidirectional user intent modeling refers to the process of modeling and predicting a user's intent or preferences in a one-way, singular direction. In this approach, the

system primarily focuses on understanding and forecasting a user's intention without considering potential changes or shifts in intent.
Unidirectional user intent modeling assumes that a user's intent remains relatively consistent during a particular interaction or session and does not actively consider the possibility of evolving preferences. It is suitable for scenarios where users have a straightforward, unchanging goal, and their intent remains stable throughout their interaction with a system.
However, this modeling approach may not capture more complex user behaviors where intent can shift or involve multiple aspects. Where a combination of long-term preferences and short-term behaviors. with traditional RNN structures have been proposed [40]. These enhancements introduce a time-aware controller and a content-aware controller to better consider contextual information for state transitions. Additionally, an attention-based framework combines long-term and short-term preferences, allowing adaptive user representation based on specific contexts. The proposed method consistently outperforms state-of-the-art techniques, offering a more effective approach to user modeling [40]. a Time-aware Long- and Short-term Attention Network (TLSAN) has been introduced [41]. TLSAN addresses the observations that users' sequential behavior records aggregate at different time positions and that users exhibit personalized preferences related to this phenomenon. It incorporates personalized time-aggregation modeling and long- and short-term feature-wise attention layers. These components enable TLSAN to adaptively utilize user preferences, handle sparse interaction data, and improve recommendation accuracy [41].

* **Self Attentive Modeling**
  Self-attentive user intent modeling is a sophisticated technique employed in recommendation systems to gain a deep understanding of user preferences and intent in a highly context-aware manner. It leverages self-attention mechanisms, commonly associated with Transformer models, to capture intricate relationships within user interactions and effectively model user intents for personalized recommendations [50]. One of the key challenges in recommendation systems is finding a balance between capturing long-term semantics and providing suitable recommendations for denser datasets. To address this, the paper introduces a novel self-attention-based sequential model called SASRec, which aims to capture long-term semantics similar to Recurrent Neural Networks but utilizes an attention mechanism to make predictions based on a small number of recent actions, akin to Markov Chains. It adapts by assigning weights to previous items at each time step, striking a balance between these two contrasting approaches. It takes inspiration from the success of the Transformer model in machine translation and adapts self-attention mechanisms to sequential recommendation problems. This approach distinguishes itself by focusing on drawing context from all past actions while framing predictions based on a limited number of recent actions. SASRec significantly adapted to data set sparsity and offered improved speed.
  In contrast, the Self-Attention Network has been known for its efficiency, parallel computing capabilities, and capacity to model dependencies. However, existing SAN-based models often struggle to distinguish users' long-term preferences from their short-term interests due to their neglect of current interests and temporal order information in sequences. To address this limitation, the paper introduces a novel multi-layer Long- and Short-term Self-Attention network designed for sequential recommendation and it divides a user's sequence into sub-sequences based on time spans and employs two

self-attention layers. The first layer captures short-term dynamics from the latest sub-sequence, while the second layer encompasses both the user's long-term preferences from previous and recent sub-sequences. These long- and short-term representations are combined into a hybrid representation
SASRec and CSAN, often fall short in fully considering the temporal order of sequences. To address this limitation, [51] introduces LSSA, which explicitly models both long-term and short-term user preferences using self-attention. LSSA splits sequences into sub-sequences, with the current session reflecting short-term interest and all previous behavior representing long-term interest. The self-attention mechanism adapts to capture the dependencies between these behaviors. Experimental results confirm the effectiveness of LSSA in sequential recommendation.
In summary, self-attentive user intent modeling, showcases the potential of self-attention mechanisms in capturing user preferences and intent in recommendation systems. These models effectively address the challenge of balancing long-term and short-term interests, thereby improving recommendation performance and providing valuable insights into user behavior.

* **Hierarchical Fusion Modeling**
Hierarchical fusion modeling for long-term and short-term user intent modeling is an advanced approach that seeks to capture and integrate different dimensions of user preferences and behaviors. It addresses the challenges of modeling evolving user interests over time, distinguishing between short-term and long-term preferences, and effectively fusing these insights to enhance recommendation systems.
Many methods just concatenates the long and short-term interests and entangles both aspects together for final prediction without explicitly distinguishing the importance of long and short-term interests to final prediction, which will lead to inferior accuracy and interpretability. To address aforementioned challenges, [60] proposed a novel method named Hierarchical Interests Fusing Network ,In the context of personalized product search [60], the Hierarchical Interests Fusing Network is introduced as an innovative solution. It recognizes the coexistence of short-term and long-term user preferences. Short-term interests are influenced by recent actions and are subject to rapid changes, while long-term interests remain relatively stable over time. Here The Short-term Interests Extractor uses three unique encoders to extract short-term interests efficiently. The Long-term Interests Extractor employs an attention mechanism to capture long-term user preferences. The Interests Fusion Module adeptly combines short-term and long-term interests, considering raw contextual features. The Interests Disentanglement Module utilizes a self-supervised framework to disentangle the intricate interplay between short-term and long-term interests. Previous approaches have unintentionally overlooked the significance of various types of historical behaviors, such as clicks, favorites, or purchases. These behaviors were often treated merely as specific features integrated into the original recommendation framework, to solve this issue [49] proposed, Hierarchical Attention Fusing Long and Short-term Preferences model is introduced as a unified framework to address the dynamic nature of users' long-term and short-term preferences. It distinguishes itself by considering both types of user preferences and employs a multi-head self-attention mechanism to capture historical behavioral patterns and multi-aspect item features. It utilizes a hierarchical attention structure to reveal item-item correlations within user behavior sequences and

effectively fuses long-term and short-term preferences. Additionally, a joint learning mechanism is introduced to combine general preferences with users' current interests. While attention mechanisms have proven effective in modeling sequential data, HAFLS goes beyond by categorizing user behavior sequences into long-term and short-term preferences, capturing general and current preferences with self-attention, and incorporating item dependencies using hierarchical attention. The outcome is a significant enhancement in sequential recommendation performance compared to existing models, positioning HAFLS as a promising approach to effectively capture and utilize long-term and short-term user intent [49].
In summary, hierarchical fusion modeling for long-term and short-term user intent modeling is a sophisticated approach that leverages various techniques and considerations to provide a comprehensive understanding of user behavior. It significantly enhances the capabilities of recommendation systems by capturing the multi-faceted and evolving nature of user preferences.

– **Bidirectional Modeling**
Bidirectional user intent modeling is an approach that comprehensively understands and predicts user preferences in both past and future interactions. Unlike unidirectional models, bidirectional modeling, exemplified by BERT4Rec, captures dependencies and patterns in user behavior by considering both the left (past) and right (future) contexts. BERT4Rec employs deep bidirectional self-attention and a Cloze objective for efficient training, resulting in enriched and context-aware user behavior representations. This approach enhances the model's ability to understand long-range dependencies and complex user intent patterns, ultimately improving recommendation performance, making it valuable for dynamic and context-aware recommendations across various domains. The problem addressed in the given context is the need to enhance recommendation accuracy, especially in content-insufficient domains, by leveraging content-rich information, user attributes, social relations, and other heterogeneous user data. Unidirectional modeling often struggle to provide accurate recommendations in such scenarios.
solution proposed in [37],U-BERT uses bidirectional modeling that customizes the BERT model for recommendation purposes. It employs a two-stage approach with distinct pre-training and fine-tuning stages. In the pre-training phase, it introduces user and review encoders. In the fine-tuning phase, it incorporates item representations and a review co-matching layer. This innovative approach improves recommendation accuracy, especially in domains with limited content data, another approach is mentioned in [52], UPRec employs BERT-like techniques to encode sequential behavior data bidirectionally. It introduces user-aware pre-training tasks like User Attribute Prediction and Social Relation Detection. These tasks help incorporate user attributes and social graphs into recommendation models, leading to more context-aware and personalized recommendations UPRec focuses on integrating heterogeneous user information, including user attributes, sequential behaviors, and social graphs. It achieves this through user-aware pre-training tasks and aligns different types of user data into a unified semantic space using a shared encoder. Here bidirectional modeling enhances the model's ability to capture long-range dependencies and understand complex patterns in user intent, ultimately leading to improved recommendation performance.

- **Multi Intent modeling**
Multi-intent modeling, in the context of recommendation systems and user profiling,

refers to the approach of modeling and predicting multiple user intentions, interests, or preferences simultaneously. Instead of assuming a single dominant intent, multi-intent modeling acknowledges that users often have diverse and concurrent intentions, which can encompass a wide range of preferences and interests. Multi-Intent User Intent Modeling is a process aimed at understanding and predicting multiple intentions or goals of a user simultaneously, particularly in the context of recommendation systems. Traditional intent modeling typically assumes a single intent for a user, but in real-world scenarios, users often have multiple concurrent intentions when interacting with a system. This limitation has prompted the introduction of a novel approach called ComiRec (Controllable Multi-Interest Recommendation), designed to address this challenge [44].ComiRec introduces a multifaceted architecture tailored for sequential recommendation tasks. At its core, it incorporates a multi-interest module capable of extracting and encompassing the diverse interests inherent in a user's behavior sequences. This module utilizes two distinct methods, the dynamic routing method and the self-attentive method, to better discern and incorporate multiple user interests. The goal is to enhance the quality and personalization of recommendations by acknowledging and effectively utilizing the multifaceted nature of user interests. [45] shows how the multi-interest user intent modeling approach plays a crucial role in the matching and ranking stages. The matching stage involves retrieving a candidate pool of items relevant to a user's interests, while the ranking stage sorts these candidates based on their alignment with user preferences. Existing deep learning models often represent users with a single vector, which can be limiting in encapsulating diverse user interests. To address this, the Multi-Interest Network with Dynamic Routing is introduced as a solution for addressing the intricacies of user interests, particularly in the matching stage of industrial recommender systems. It incorporates a multi-interest extractor layer that employs dynamic routing to generate multiple user representation vectors, each corresponding to various facets of user interests. These vectors collectively paint a nuanced picture of a user's diverse interests and can be effectively utilized in the matching stage, even when dealing with extensive item pools.

2.3.2 Intent Modeling in Conversational Multiround Recommender.

User intent modeling in conversational recommendation systems is the process of comprehending and forecasting users' intentions and preferences during interactions with a system, with the ultimate objective of furnishing tailored recommendations or responses. This modeling is instrumental in crafting context-sensitive and user-centric conversational engagements. Let's delve into the fundamental facets and strategies pertaining to user intent modeling within these systems:

Navigation-by-Preference (n-by-p) is an illustrative example of a conversational recommender system introduced in [34]. It harnesses preference-based feedback to aid users in traversing item space, with a focus on uncovering items that align with both their enduring preferences, cataloged in their user profile, and their momentary preferences, inferred from feedback during the ongoing conversation. This adaptability renders it well-suited for domains characterized by scant structured item descriptions, such as music and movies. Notably, it offers configurational flexibility, enabling it to disregard long-term preferences and concentrate solely on positive feedback or consider a spectrum of both positive and negative feedback. Moreover, it can factor in preceding feedback rounds or rely solely on the most recent feedback, seamlessly amalgamating preference-based feedback and user profiles.

The prevalent challenge lies in users harboring diverse interests spanning various categories, a complexity often underestimated by presuming that a single user interest can be encapsulated by

closely related items. User-Centric Conversational Recommendation (UCCR), as outlined in [35], provides a remedy for this predicament. UCCR strives to systematically model multi-aspect user preferences within Conversational Recommendation Systems. This is achieved by fortifying user-centric preference learning through the assimilation of users' prevailing session preferences and historical session preferences, while also taking cognizance of the preferences of akin users. UCCR adeptly captures the manifold facets of users' interests through the acquisition of multi-aspect user preferences.

Furthermore, an alternative approach is presented in [36], which tackles the conundrum of modeling multiple user interests. This method harnesses hierarchical knowledge of items and categories to shape a user profile replete with diverse interests across varying levels of granularity. By capitalizing on this hierarchical framework, the model accommodates multifarious interests spanning different dimensions. Notably, it explicitly accounts for multiple abstract item categories extracted from the hierarchical entity knowledge to generate user portrait vectors. This furnishes a more holistic representation of the user's interests, culminating in recommendations that are not only more pertinent but also more personalized.

In sum, user intent modeling in conversational recommendation systems is indispensable for delivering tailored recommendations and responses. Examples like Navigation-by-Preference (n-by-p), User-Centric Conversational Recommendation (UCCR), and the approach delineated in [36] exemplify diverse strategies for addressing the challenges posed by users' multifaceted interests and preferences. These approaches underscore the significance of incorporating multi-aspect user preferences and hierarchical knowledge to enhance the precision and personalization of recommendations.

3 USER INTENT MODELING USE CASES IN DIFFERENT DOMAINS

User intent modeling is valuable in various domains and can have a significant impact on improving user experiences and outcomes. Here are some use cases of user intent modeling in different domains

3.1 News

In the field of news recommendation, accurately modeling user intent is crucial for delivering personalized content. Traditional methods often rely on static historical data, which may not capture the evolving nature of user preferences over time and the dynamic nature of news. This research addresses the challenge of evolving user interests in news recommendations and proposes a recommendation approach that considers both long-term and short-term user profiles to emphasize diversity in recommendations [53]. By integrating both long-term and short-term user reading preferences, this method recognizes the stability of users' overall interests alongside their changing specific interests. It differentiates newsgroups based on long-term profiles and selects news articles within those groups using the user's short-term profile, providing relevant and up-to-date news recommendations. Additionally, to enhance recommendation diversity, an absorbing random walk model is applied to a user-item affinity graph, broadening users' preferences by introducing diverse topics.

Another aspect of modeling user intent in news recommendation involves focusing on users' high-level, goal-oriented reading intentions, which significantly influence their news preferences [54]. The paper introduces the intention-aware personalized news recommendation model to address this challenge. It aims to model a user's reading intentions and preferences for personalized news recommendations, going beyond preferences to capture users' intrinsic reading intentions and transitions over time. This comprehensive framework includes modules for detecting intentions,

modeling transitions, and capturing preferences. These modules are integrated to create an intention-aware user representation, and a prediction module estimates the click probability for each candidate news article in the user's next click action. By providing a comprehensive approach to accurately model users' reading intentions alongside their preferences in personalized news recommendations, contributes to more accurate and personalized news recommendations [54].

3.2 POI

In the domain of location-based social networks, Point-of-Interest (POI) recommendations are essential for predicting a user's next destination based on their check-in patterns. Previous research often overlooked incorporating spatiotemporal data and user preferences. To address this gap, the Long- and Short-Term Preference Modeling based on the Multi-level Attention framework (LSMA) was introduced, aiming to capture both long-term and short-term user preferences while considering spatio-temporal information. It also addresses the periodic nature of user check-ins and explores user interests comprehensively, resulting in superior performance compared to existing recommendation systems [46].While previous attempts in POI recommendation included Markov chains and recurrent neural networks, such as LSTM, they struggled to incorporate critical contextual data. LSMA addresses these limitations by comprehensively harnessing check-in information, considering both long- and short-term preferences, check-in periodicity, and category transitions, improving recommendation accuracy [46]. LSTPM utilizes a nonlocal network to capture long-term preferences and a geo-dilated RNN for short-term preferences, resulting in more reliable recommendations [55]. POIFormer disentangles mobility patterns and user preferences, employing components for mobility pattern encoding, user preference generation, and preference decoding to enhance the accuracy of next location recommendations [56]. LSMA is employed to address issues related to spatio-temporal data, user check-in regularity, and category preferences in POI recommendations, consistently outperforming existing systems [57]. UTSR introduces a novel sequential recommendation model that combines user preferences and spatiotemporal information to enhance recommendation quality and accuracy [58]. Finally, RTPM focuses on real-time POI recommendations, considering both long-term and short-term user preferences and category filtering while prioritizing user privacy, ultimately improving recommendation accuracy [59]. These research endeavors collectively underscore the significance of modeling long-term and short-term user preferences, leveraging spatio-temporal data, and disentangling various factors to enhance the accuracy of location-based recommendations, providing improved systems for diverse user contexts.

3.3 Ecommerce

In the e-commerce sector, understanding user intent is crucial for providing personalized product recommendations. By analyzing a user's browsing and purchase history, as well as their search queries, e-commerce platforms can model user intent to suggest relevant products, resulting in increased sales and customer satisfaction In the context of e-commerce platforms, understanding and predicting user behavior, which encompasses long-term interests and short-term needs, is a complex task. Traditional methods like Markov chains, CNNs, and RNNs have limitations in comprehensively capturing this dynamic mix of characteristics. To address this challenge, the paper introduces an attention-based deep neural network (ADNNet). ADNNet combines the strengths of CNNs for short-term patterns and GRUs for long-term patterns in user behavior sequences. It distinguishes itself through the inclusion of an attention mechanism, dynamically adapting to the specific dynamics of each user's behavior, which enhances its effectiveness in capturing user behavior influenced by both immediate needs and persistent interests in the e-commerce context [6].

Understanding and modeling user preferences in e-commerce, particularly in the context of personalized product search, is a formidable challenge as user preferences encompass both long-term and short-term dimensions, significantly influencing purchasing decisions. To address this limitation, the Attentive Long Short-Term Preference (ALSTP) [39]model has been introduced, effectively integrating both long-term and short-term user preferences with the current search query through two attention networks, resulting in more precise modeling of users' search intentions. In the context of e-commerce interactions, driven by distinct user intents such as clicks, cart additions, or purchases, the enhancement of recommendation systems necessitates capturing both short-term and long-term user interests[42], as well as discovering latent intent. Hierarchical Interests Fusing Network (HIFN) offers an innovative approach to capture users' short-term and long-term interests effectively [60], addressing challenges related to interest extraction, integration, and disentanglement through its four key modules, excels in capturing user behavior dynamics and preferences, even in scenarios with less explicit user intent, effectively resolving the sparsity issue inherent in e-commerce datasets by focusing on interactions at the category level. Moreover, in the context of Click-Through Rate estimation for personalized product search, the comprehensive framework of the Hierarchical Interests Fusing Network introduces four pivotal modules that collectively offer a groundbreaking approach to elevate CTR prediction in personalized product search by adeptly extracting and integrating both short-term and long-term user interests while navigating the complexities of user behavior and preferences, effectively addressing the challenges in this domain.

3.4 Music

In the study presented in [61], the focus is on improving playlist generation for music streaming platforms, with an emphasis on incorporating both short-term and long-term user preferences. While existing algorithms typically prioritize recent listening history and context for playlist creation, they tend to overlook users' enduring music preferences and social network connections. To enhance playlist personalization, various strategies are evaluated, considering multi-dimensional user-specific preference signals like liked tracks, favorite artists, semantically similar tracks, co-occurring tracks, and tracks appreciated by the user's social network friends. These signals prove crucial in enhancing playlist personalization quality. The research showcases that accounting for long-term preferences, including track co-occurrence patterns and favorite track repetitions, significantly improves playlist accuracy, measured through the track hit rate (recall). Additionally, the study evaluates the impact of personalization strategies on recommendation diversity and coherence with recent listening history, aiming to provide more relevant and diverse music recommendations to users, thus contributing to the field of music recommendation.

Moving to [62], the research addresses the challenges of recommending tracks in online music streaming services, acknowledging the unique characteristics of the music domain, including track length, context dependency, and session-based consumption. The proposed CoSeRNN neural network architecture aims to model users' preferences at the beginning of each session, utilizing vector-space embeddings of tracks. The model leverages recurrent neural networks to combine long-term, context-independent preferences with session-specific and context-dependent offsets, generating session-based embeddings. Experimental evaluations consistently show that CoSeRNN outperforms other baseline methods in session ranking and track ranking tasks. The study highlights the model's effectiveness in diverse contexts and underscores the importance of integrating both sequential and contextual information for accurate music recommendations. In summary, [62] introduces CoSeRNN as a promising solution to the challenges of music recommendation in online streaming services, enhancing user satisfaction across various sessions and advancing the field of music recommender systems.

4 CONCLUSION

The findings from this review contribute to a deeper understanding of user intent modeling and its critical role in shaping the next generation of personalized recommendation systems. As the research landscape evolves, this paper is a foundational resource to guide future research and development in pursuing more effective and user-centric recommendation algorithms.By effectively modeling user intent, recommendation systems can better anticipate user needs, adapt to dynamic preferences, and deliver more meaningful and context-aware recommendations. This paper presents an overview of the state-of-the-art methods and analyzes their strengths and limitations in modelling user intent.

REFERENCES

[1] Chen, Wanyu, et al. "Improving end-to-end sequential recommendations with intent-aware diversification." Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020.

[2] Wasilewski, Jacek, and Neil Hurley. "Intent-aware diversification using a constrained PLSA." Proceedings of the 10th ACM Conference on Recommender Systems. 2016.

[3] S. Vargas, P. Castells, and D. Vallet. Intent-oriented Diversity in Recommender Systems. ACM SIGIR'11 Conference Proceedings, pages 1211–1212, 2011.

[4] Kaya, Mesut, and Derek G. Bridge. "Intent-Aware Diversification using Item-Based SubProfiles." RecSys Posters. 2017.

[5] Kaya, Mesut, and Derek Bridge. "Subprofile-aware diversification of recommendations." User Modeling and User-Adapted Interaction 29 (2019): 661-700.

[6] Yan, Cairong, et al. "Modeling Long-and short-term user behaviors for sequential recommendation with deep neural networks." 2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021.

[7] Zhou, Lei, et al. "Personalized recommendation algorithm based on user preference and user profile." Web, Artificial Intelligence and Network Applications: Proceedings of the Workshops of the 34th International Conference on Advanced Information Networking and Applications (WAINA-2020). Springer International Publishing, 2020

[8] Yin, Hongzhi, et al. "Modeling location-based user rating profiles for personalized recommendation." ACM Transactions on Knowledge Discovery from Data (TKDD) 9.3 (2015): 1-41.

[9] Lian, Jianxun, et al. "Multi-interest-aware user modeling for large-scale sequential recommendations." arXiv preprint arXiv:2102.09211 (2021).

[10] Zhou, Wen, and Wenbo Han. "Personalized recommendation via user preference matching." Information Processing & Management 56.3 (2019): 955-968.

[11] Kaya, Mesut, and Derek Bridge. "Accurate and diverse recommendations using item-based subprofiles." The Thirty-First International Flairs Conference. 2018.

[12] Chang, Bo, et al. "Latent User Intent Modeling for Sequential Recommenders." Companion Proceedings of the ACM Web Conference 2023. 2023.

[13] Zanitti, Michele, Sokol Kosta, and Jannick Sørensen. "A user-centric diversity by design recommender system for the movie application domain." Companion Proceedings of the The Web Conference 2018. 2018.

[14] Kim, Hyoung-Rae, and Philip K. Chan. "Learning implicit user interest hierarchy for context in personalization." Applied Intelligence 28 (2008): 153-166

[15] Widyantoro, D.H., Ioerger, T., Yen, J.: Learning User Interest Dynamics with a ThreeDescriptor Representation. Journal of the American Society for Information Science and Technology, Vol. 52 (2001) 212-225

[16] Chen, C.C., Chen, M.C., Sun, Y.: PVA: A Self-Adaptive Personal View Agent. Journal of Intelligent Information Systems, Vol.18 (2002) 173-194

[17] Basile, Pierpaolo, et al. "Modeling Short-Term Preferences in Time-Aware Recommender Systems." UMAP Workshops. 2015.

[18] Quadrana, Massimo, et al. "Personalizing session-based recommendations with hierarchical recurrent neural networks." proceedings of the Eleventh ACM Conference on Recommender Systems. 2017.

[19] Hidasi, Balázs, et al. "Session-based recommendations with recurrent neural networks." arXiv preprint arXiv:1511.06939 (2015).

[20] Ruocco, Massimiliano, Ole Steinar Lillestøl Skrede, and Helge Langseth. "Inter-session modeling for session-based recommendation." Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems. 2017.

[21] Yeganegi, Reza, and Saman Haratizadeh. "STAR: A Session-Based Time-Aware Recommender System." arXiv preprint arXiv:2211.06394 (2022).

[22] Wang, Ruiqin, Jungang Lou, and Yunliang Jiang. "Session-based recommendation with time-aware neural attention network." Expert Systems with Applications 210 (2022): 118395.

[23] Kompan, Michal, Ondrej Kassak, and Maria Bielikova. "The short-term user modeling for predictive applications." Journal on Data Semantics 8 (2019): 21-37.
[24] Guo, Jiayan, et al. "Learning multi-granularity consecutive user intent unit for session-based recommendation." Proceedings of the fifteenth ACM International conference on web search and data mining. 2022.
[25] Oh, Sejoon, et al. "Implicit session contexts for next-item recommendations." Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022.
[26] Zhang, Xiaokun, et al. "Price does matter! modeling price and interest preferences in session-based recommendation." Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2022.
[27] Wu, Shu, et al. "Session-based recommendation with graph neural networks." Proceedings of the AAAI conference on artificial intelligence. Vol. 33. No. 01. 2019.
[28] Wang, Ting-Yun, et al. "Modeling inter-session information with multi-interest graph neural networks for the next-item recommendation." ACM Transactions on Knowledge Discovery from Data 17.1 (2023): 1-28.
[29] Qiu, Ruihong, et al. "Exploiting inter-session information for session-based recommendation with graph neural networks." ACM Transactions on Information Systems (TOIS) 38.3 (2020): 1-23.
[30] Cao, Wenming, et al. "Implicit user relationships across sessions enhanced graph for session-based recommendation." Information Sciences 609 (2022): 1-14.
[31] Cai, Chenwei, Ruining He, and Julian McAuley. "SPMC: Socially-aware personalized Markov chains for sparse sequential recommendation." arXiv preprint arXiv:1708.04497 (2017).
[32] Rendle, Steffen, Christoph Freudenthaler, and Lars Schmidt-Thieme. "Factorizing personalized markov chains for next-basket recommendation." Proceedings of the 19th international conference on World wide web. 2010.
[33] He, Ruining, and Julian McAuley. "Fusing similarity models with markov chains for sparse sequential recommendation." 2016 IEEE 16th international conference on data mining (ICDM). IEEE, 2016.
[34] Rana, Arpit, and Derek Bridge. "Navigation-by-preference: a new conversational recommender with preference-based feedback." Proceedings of the 25th International Conference on Intelligent User Interfaces. 2020.
[35] Li, Shuokai, et al. "User-centric conversational recommendation with multi-aspect user modeling." Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2022.3
[36] Okuda, Yuka, et al. "Modeling Multiple User Interests using Hierarchical Knowledge for Conversational Recommender System." arXiv preprint arXiv:2303.00311 (2023).
[37] Qiu, Zhaopeng, et al. "U-BERT: Pre-training user representations for improved recommendation." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 5. 2021.
[38] Zhu, Yu, et al. "What to Do Next: Modeling User Behaviors by Time-LSTM." IJCAI. Vol. 17. 2017.
[39] Guo, Yangyang, et al. "Attentive long short-term preference modeling for personalized product search." ACM Transactions on Information Systems (TOIS) 37.2 (2019): 1-27.
[40] Yu, Zeping, et al. "Adaptive User Modeling with Long and Short-Term Preferences for Personalized Recommendation." IJCAI. 2019.
[41] Zhang, Jianqing, Dongjing Wang, and Dongjin Yu. "TLSAN: Time-aware long-and short-term attention network for next-item recommendation." Neurocomputing 441 (2021): 179-191.
[42] Tanjim, Md Mehrab, et al. "Attentive sequential models of latent intent for next item recommendation." Proceedings of The Web Conference 2020. 2020.
[43] Zhu, Nengjun, et al. "Learning a hierarchical intent model for next-item recommendation." ACM Transactions on Information Systems (TOIS) 40.2 (2021): 1-28.
[44] Cen, Yukuo, et al. "Controllable multi-interest framework for recommendation." Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020.
[45] Li, Chao, et al. "Multi-interest network with dynamic routing for recommendation at Tmall." Proceedings of the 28th ACM international conference on information and knowledge management. 2019.
[46] Wang, Xueying, et al. "Long-and Short-Term Preference Modeling Based on Multi-Level Attention for Next POI Recommendation." ISPRS International Journal of Geo-Information 11.6 (2022): 323.
[47] Neil, Daniel, Michael Pfeiffer, and Shih-Chii Liu. "Phased lstm: Accelerating recurrent network training for long or event-based sequences." Advances in neural information processing systems 29 (2016).
[48] Sun, Fei, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. "BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer." In Proceedings of the 28th ACM international conference on information and knowledge management, pp. 1441-1450. 2019.
[49] Du, Yongping, et al. "A unified hierarchical attention framework for sequential recommendation by fusing long and short-term preferences." Expert Systems with Applications 201 (2022): 117102.
[50] Kang, Wang-Cheng, and Julian McAuley. "Self-attentive sequential recommendation." 2018 IEEE international conference on data mining (ICDM). IEEE, 2018.

[51] Xu, Chengfeng, et al. "Long-and short-term self-attention network for sequential recommendation." Neurocomputing 423 (2021): 580-589.
[52] Xiao, Chaojun, et al. "UPRec: User-Aware Pre-training for Recommender Systems. arXiv preprint (2021)." (2021).
[53] Li, Lei, et al. "Modeling and broadening temporal user interest in personalized news recommendation." Expert Systems with Applications 41.7 (2014): 3168-3177.
[54] Wang, Rongyao, et al. "Intention-Aware User Modeling for Personalized News Recommendation." International Conference on Database Systems for Advanced Applications. Cham: Springer Nature Switzerland, 2023.
[55] Sun, Ke, et al. "Where to go next: Modeling long-and short-term user preferences for point-of-interest recommendation." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 01. 2020.
[56] Luo, Yan, et al. "End-to-End Personalized Next Location Recommendation via Contrastive User Preference Modeling." arXiv preprint arXiv:2303.12507 (2023).
[57] Wang, Xueying, et al. "Long-and Short-Term Preference Modeling Based on Multi-Level Attention for Next POI Recommendation." ISPRS International Journal of Geo-Information 11.6 (2022): 323.
[58] Yin, Sizhe, et al. "Fusing User Preferences and Spatiotemporal Information for Sequential Recommendation." IEEE Access 10 (2022): 89545-89554.
[59] Liu, Xin, et al. "Real-time POI recommendation via modeling long-and short-term user preferences." Neurocomputing 467 (2022): 454-464.
[60] Shen, Qijie, et al. "Hierarchically Fusing Long and Short-Term User Interests for Click-Through Rate1 Prediction in Product Search." Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022.
[61] Kamehkhosh, Iman, Dietmar Jannach, and Lukas Lerche. "Personalized Next-Track Music Recommendation with Multi-dimensional Long-Term Preference Signals." UMAP (Extended Proceedings). 2016.
[62] Hansen, Casper, et al. "Contextual and sequential user embeddings for large-scale music recommendation." Proceedings of the 14th ACM Conference on Recommender Systems. 2020.