Micro-Video Recommendation Systems: Overview of Techniques, Challenges, and Advancements

MIGUEL MATOS, Departamento de Eletrónica, Telecomunicações e Informática, Portugal

This work provides an overview of the latest advancements in micro-video recommendation systems, delving into the research that aims to push the boundaries of the field. A summary of the state-of-the-art methodologies is presented, followed by three recent works that refine these traditional methods through different methods. These papers demonstrate how current models are starting to employ specialized and behavior-focused techniques, seeking refinement of current methods.

CCS Concepts: • **Information systems** → **Recommendation Systems**.

Additional Key Words and Phrases: Micro-Video, Recommendation Systems

ACM Reference Format:

Miguel Matos. 2024. Micro-Video Recommendation Systems: Overview of Techniques, Challenges, and Advancements. *J. ACM* 37, 4, Article 111 (August 2024), 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

In the digital era, the rapid proliferation of online video content has revolutionized how we consume media, marking a significant shift from traditional broadcasting to on-demand, user-centric platforms. This evolution has birthed the necessity for sophisticated video recommendation systems, vital tools in navigating the vast ocean of digital content. These systems, leveraging advanced algorithms and user data, curate personalized video content, enhancing user experience and engagement. Their importance is magnified in the context of the latest trends in micro-videos – short and concise video segments that have gained immense popularity in social media landscapes.

Historically, the journey of video recommendation systems can be traced back to the early days of online video platforms. Initially, these systems were rudimentary, relying on simple algorithms based on popularity metrics or basic user preferences. The advent of platforms like YouTube marked a significant milestone, introducing more complex algorithms that considered factors such as viewing history, user ratings, and demographic data [1–4].

As technology advanced, so did these systems. The introduction of machine learning and artificial intelligence ushered in an era of predictive analytics and personalization at an unprecedented scale. Most modern systems employ deep learning techniques to enhance the recommendation process, as seen in [3], [2], and [1].

The recent surge in micro-video content, predominantly on platforms like TikTok, Instagram Reels, and Snapchat, has further nuanced the requirements of recommendation systems. These micro-videos, often lasting just a few seconds to a minute, cater to the rapidly diminishing attention spans of modern digital consumers and their preference for quick, engaging content. The transient nature of these videos demands that recommendation systems not only be accurate but also extremely responsive to the ever-changing trends and user preferences.

Author's address: Miguel Matos, miguelamatos@ua.pt, Departamento de Eletrónica, Telecomunicações e Informática, Aveiro, Portugal.

Manuscript submitted to ACM

In this monograph, we delve into the state-of-the-art in micro-video recommendation systems. We explore three inventive research papers that push the boundaries of current technology and shine a light on how modern recommendation systems are aiming to tailor their predictions by modeling deeper and non-explicit user behavior.

2 VIDEO AND MICRO-VIDEO RECOMMENDATION SYSTEMS

The challenge of recommending a video from an uncountable amount of them can also be seen as the task of filtering the bad ones. The work pursued at [4] provides an overview of the classification and state-of-art recommendation systems. The most defining methodologies that shape these systems include [4] collaborative filtering, content-based filtering, and hybrid approaches.

2.1 Content-Based Filtering

This method employs an examination of the intrinsic attributes or features of videos. Such features encompass a broad spectrum, including the video's genre, tags, and description, alongside its visual and auditory content, and any associated metadata.

Moreover, these systems are designed to construct a detailed user profile, which is derived from the individual's historical interactions with various items. This can encapsulate the user's history of watched, liked, or shared videos, as well as the characteristic features of these videos.

The recommendation process within these systems can be summed up in two parts [4]: feature extraction and profile matching. At the feature extraction level, the system extracts and analyzes features from the videos. Advanced techniques like natural language processing for text data, image and video analysis using deep learning for visual content, and audio analysis for soundtracks can be employed. At the profile matching level, the system matches the features of new or unwatched videos with the previously built user's profile and recommends videos with the closest match.

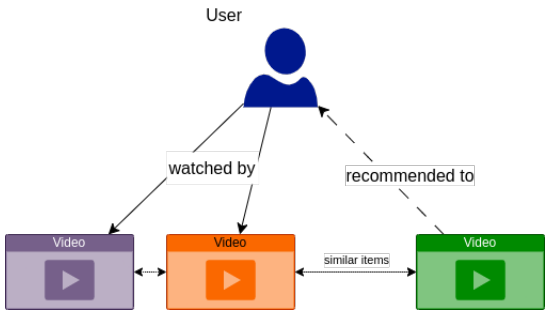

Fig. 1. High-level depiction of the recommendation process of a content-based filtering method.

2.1.1 Challenges. This approach relies heavily on the ability to analyze the content it recommends. If the analysis of such content does not provide meaningful results, due to, for example, poor video feature extraction, the recommendation process will suffer and might provide poor or irrelevant suggestions. Also, due to only taking into account information about the user's historical preferences, it might miss out on dynamic interest, as in interests that change over time. This can lead to a "filter bubble", which only recommends content similar to what the user has interacted with in the past [4].

Recommending items to users of which there is no profile information available, i.e. new users, can also be challenging since there is no previous information to use to shape the recommendations made by the system. This is also denominated as the cold-start problem [2–4].

2.2 Collaborative Filtering

Collaborative Filtering (CF) primarily operates on a matrix of user-item interactions. These interactions could be explicit (like ratings) or implicit (like views or likes on a micro-video). In contrast with content-based filtering, this method considers information about other similar users to predict what a user would like.

This approach can be split into two main methodologies [4]. User-based CF finds users similar to the target user and recommends items similar to the ones those users have liked. Similarity is typically measured using metrics like cosine similarity or Pearson correlation. On the other hand, Item-based CF finds items similar to those that the user has already liked or interacted with, instead of finding similar users to the target user. It's often computationally more efficient than user-based filtering, especially for systems with more items than users. It is also important to note that, contrary to content-based filtering, this approach deems two videos as similar if they were rated or interacted with in similar ways by different users, instead of looking at the actual content and features of the multimedia, which makes it fall under the collaborative filtering category.

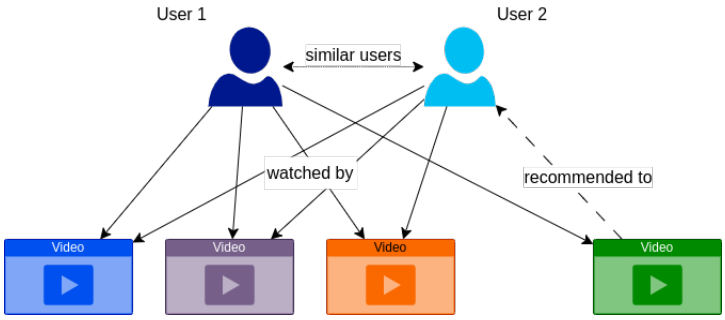

Fig. 2. High-level depiction of the recommendation process of a user-based collaborative filtering method.

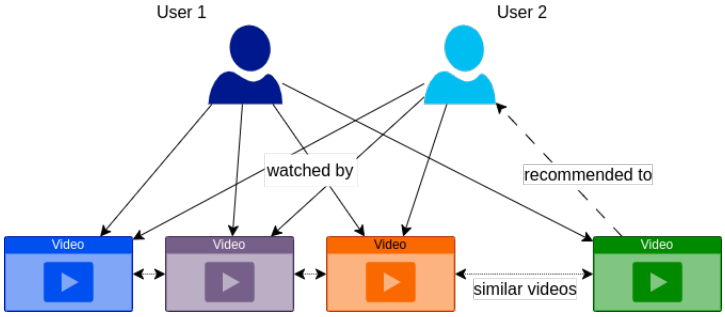

Fig. 3. High-level depiction of the recommendation process of an item-based collaborative filtering method.

overview

157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208

2.2.1 Challenges. This approach also suffers from the cold-start problem since it bases its recommendations on user-item interactions, and new users and new items do not have interaction data that can be used as information to feed the recommendation process [3, 4].

Furthermore, these methods tend to suffer from popularity bias. Since the recommendation process is mainly based on the user-item interaction matrix, items with more interactions or ratings are often more valued than less popular items, with no correlation to their relevance [4].

Data sparsity is also an aspect that can greatly influence the performance of such methods. In many systems, and especially in micro-video, the number of items greatly exceeds the number of items a user can rate or interact with. This leads to a sparse user-items matrix, which makes it difficult to find similar items and/or users, which worsens the quality of the recommendations [3, 4].

Due to the vast amount of people and items on the web, the amount of processing resources it takes to determine how similar two users are to provide meaningful suggestions is not scalable. Thus, scalability also presents an issue for this type of system [4].

2.3 Hybrid Approaches

Incorporating a hybrid strategy that combines the strengths of content-based and collaborative filtering approaches can effectively address the drawbacks of each method. This approach, known as hybrid filtering, leverages both item feature information (content-based filtering) and user interaction data (collaborative filtering) to generate recommendations that are relevant and diverse [4].

Content-based filtering excels in providing recommendations highly tailored to individual user preferences by analyzing item features. However, this method often falls short in capturing the diversity of user interests, as it primarily relies on historical user data, leading to a narrow focus on past interests. On the other hand, collaborative filtering harnesses collective user behavior to offer recommendations. While this method is effective in capturing popular trends and communal preferences, it faces challenges like the cold-start problem, where new users or items with limited interaction data are difficult to recommend accurately. It also tends to have a popularity bias, where commonly liked items are frequently recommended, potentially overlooking niche or less popular items.

By integrating these two methods, the hybrid strategy can mitigate these limitations. For new users or items (cold-start problem), the system can use content-based features to provide initial recommendations, thus enabling the collaborative filtering system to gradually learn from accumulating interaction data. In addressing the popularity bias of collaborative filtering, the hybrid approach ensures a diverse range of recommendations by incorporating unique item features from the content-based method, thus recommending items that may not be popular but align with the user's specific preferences.

Hybrid models can also adapt the balance between content-based and collaborative signals based on data availability and reliability. In scenarios with rich user interaction data, they can prioritize collaborative filtering to leverage broader user behavior patterns. Conversely, with limited interaction data, these systems can rely more on content-based features.

Overall, the hybrid filtering approach combines the detailed analysis of user profiles from content-based filtering with the broad insight into user behavior patterns offered by collaborative filtering. This results in a more comprehensive recommendation system that not only caters to individual preferences but also introduces diverse and potentially undiscovered content, effectively balancing personalization with discovery.

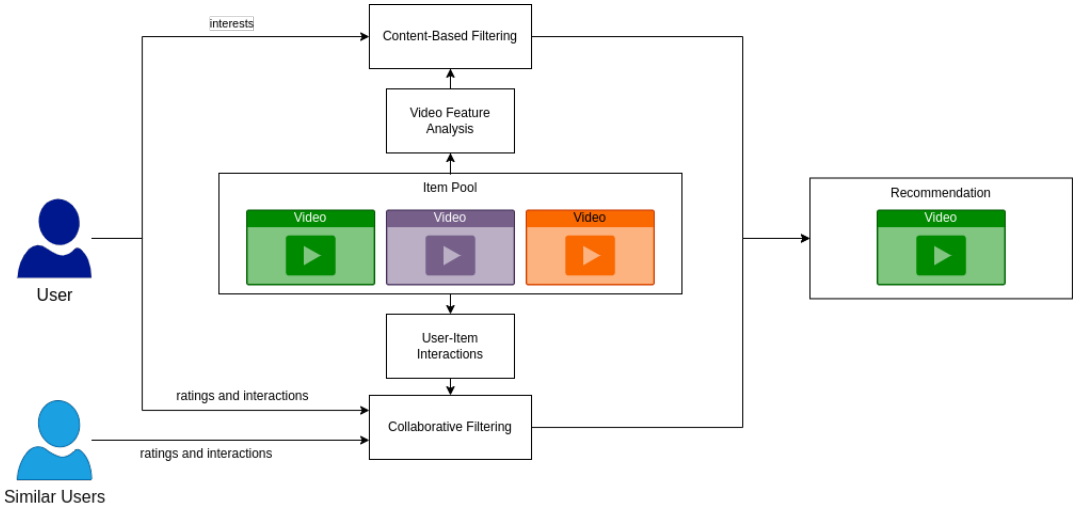

Fig. 4. High-level depiction of the recommendation process of an item-based collaborative filtering method.

2.4 Micro-Video Nuances

Micro-videos, by definition, differ from regular videos, which forces recommendation systems tailored for this shorter content to face nuances specific to these.

Micro-videos are usually shorter, which leads to a vast volume of content being generated and consumed. This leads to bigger interaction graphs, and since a user won't interact with all micro-videos they watch during a single session, sparser too, as shown in the datasets used for experimentation in the studied works. To illustrate, the experiments conducted in [1] go over two datasets with sparsity percentages of 99.77% and 99.64%. This lack of interaction also implies that user feedback for most micro-videos watched is not explicit, like a like or follow, but implicit, such as watching until the end or skipping the video in the earlier clips.

3 EMERGING TRENDS AND TECHNIQUES

In this swiftly evolving domain, recent works have been demonstrating inventive approaches to better curate content for users. This section delves into four research papers, each presenting a unique and innovative to enhance micro-video recommendation systems.

3.1 Fine-Grained User Interests

Shang et al. [3] propose a framework called FRAME whose novelty primarily focuses on processing individual clips within videos to understand user preferences more accurately. Unlike traditional methods that analyze videos as a whole, this paper aims to better understand more granular user interests by integrating clip-level feedback into the system by analyzing the feedback of a user on a micro-video instance throughout the sequence of each of its segments - in other words, clips.

The framework can be decomposed into four layers: the visual-enhanced embedding layer, the fine-grained user-preference graph convolutional layer, the preference-fusion based prediction layer, and the hybrid supervision learning layer.

Visual-Enhanced Embedding Layer. This layer uses a convolutional neural network to extract the visual features of clips in each video. The authors claim that the content of each clip on a micro-video can vary substantially, and thus the clip-level feature extraction provides useful information for the recommendation process. Before feeding these features to the next layer, they are transformed into embeddings in the user-preference space, since not all features of a video are useful for preference learning. Aspects such as the background of a scene are removed from the information within the embedding so that the information within it can be of more relevance to the overall process.

Fine-Grained User-Preference Graph Convolutional Layer. Based on the previous embeddings of video clips, the user-clip relations are built, of which there are two kinds: skip (negative) and non-skip (positive). Negative relations are restricted to clips the user did skip the video on and do not cover non-watched clips. A graph convolutional layer is then used to aggregate users' positive and negative interests.

Preference-Fusion Based Prediction Layer. Using the signals processed before, the system constructs detailed user profiles. It utilizes a multi-layer perception network to make clip-level predictions for the interest of a given user. Positive and negative preferences are adaptively fused, which means that the weighting of positive and negative feedback can be dynamically adjusted based on the specific characteristics of the user and clips under consideration.

Hybrid Supervision Learning. The model further refines its recommendations through the use of a hybrid loss function. One of the components works at a user-clip level (user-clip pointwise loss), and the other at a clip-pair level (clip-pair pairwise loss). The first aims to assess the model's ability to accurately predict interaction, and the other is meant to ensure that the model not only predicts whether a user will like a clip but also correctly ranks multiple clips in the order of the user's preference. This loss function acts as a guide for the model during the training process, indicating how well the model is performing.

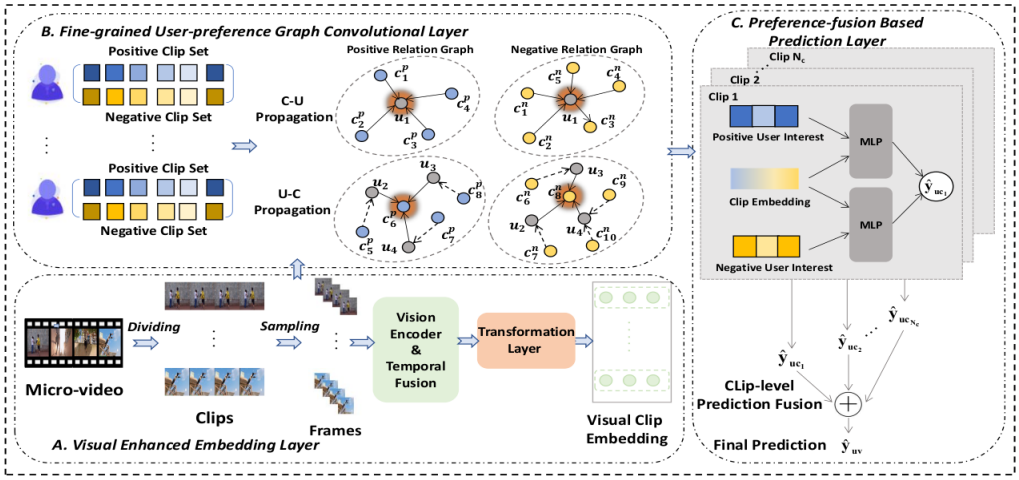

Fig. 5. Architecture of the FRAME model.

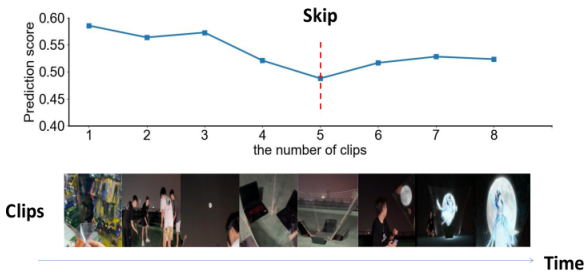

Fig. 6. Visualization on the output of the FRAME model.

Figure 5 depicts a more detailed overview of the architecture of the FRAME framework. Figure 6 shows a visualization of an example output of this framework, from one of the conducted experiments. The blue line depicts the predicted interest score over each clip of the video. The red line signals the clip at which the user skipped the video, which the model deemed the less interesting. Following the reasoning that users will keep watching a video until the content makes them intolerable or bored, the effectiveness of integrating user feedback at clip-level is demonstrated.

3.2 Multi-Trend Prediction

The DMR framework, proposed by Yujie Lu et al. [2], has a focus more shifted toward the dynamism of user interests. The authors claim that user preferences and interests are not static but dynamic, and change over time. Thus, current models that base their predictions solely on logged data might be out-of-date, or at least limited to the interaction history. This framework aims to predict how users' interests will evolve by analyzing patterns contained within user activity that might indicate shifting user preferences or emerging interests.

The framework is split into three main modules: the Pearson Correlation Coefficient enhanced implicit user network module, the history-future multi-trend joint routing module, and the multi-level time-aware attention module.

Implicit User Network. This module constructs a network based on users' interactions with micro-videos. It selects users with similar micro-video preferences and extracts future sequences of micro-videos based on their behaviors. The idea is to capture and represent the dynamic preferences and trends of users, even beyond their historical behaviors. This aims to enhance the system's understanding of diverse user interests and improve recommendation quality and also lays the foundation for capturing trends in user preferences.

Multi-trend Routing. This module focuses on grouping micro-videos from a user's historical and future sequences into diverse trends. By doing so, it aims to capture trend information from both sets of sequences, allowing for a better understanding of the evolving user preferences over time. Ultimately, it enables the system to represent the diverse trends and preferences of users, enhancing the recommendation process.

Multi-level Time Attention Mechanism. This module employs a multi-level time-aware attention mechanism to generate representations of historical and future trends. It aims to capture the temporal dynamics of user preferences by weighing the influence of historical actions on current action, considering not only what action took place but also when and how it occurred. This enhances the system's understanding of the temporal aspects of user preferences and leads to improved recommendations.

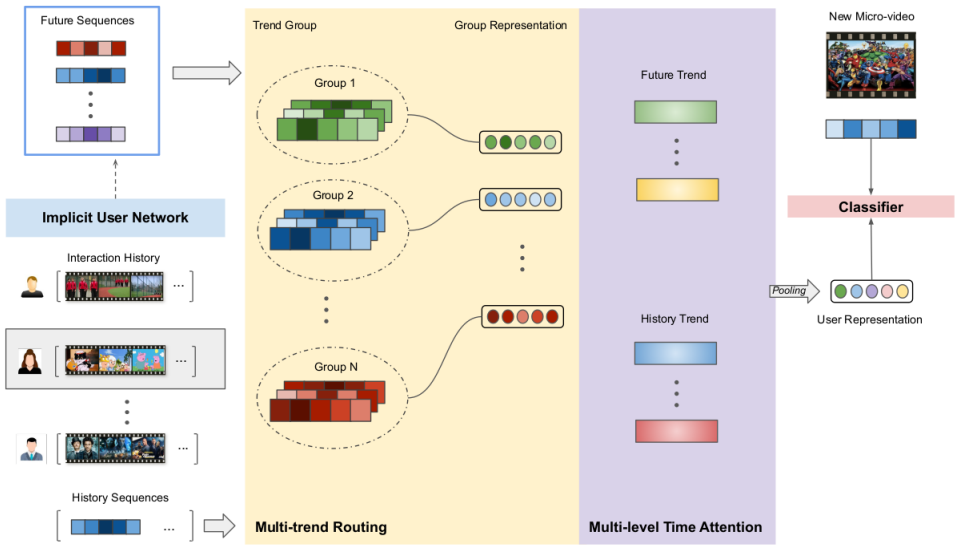

Fig. 7. Architecture of the DRM model.

Overall, this model aims to provide better recommendations by integrating over-time patterns in user behavior, which can be seen as trends. It also keeps updating its self-built user profiles with the latest trend information, to keep predictions relevant over time. Figure 7 illustrates the overview of the DMR model architecture.

3.3 Multi-Modal Recommendation

This approach is focused on learning user preferences throughout the different modalities of multimedia videos, such as visual, audio, and text, and on exploring users' deeper preferences over these different modalities.

The following paragraphs highlight the general recommendation process of this framework.

User-Item Interaction Graph. A user-item interaction graph is constructed to learn the representation of each node from interaction data and to capture higher-order collaborative signals between users and items. An existing graph convolutional network is employed to propagate user and item information through graph nodes to obtain the embeddings of users and items. This also captures high-order collaborative signals for the user-item relation.

Item-Item Modal Similarity Graph. Since the user-item graph only conveys user-item interaction information, the model also mines the potential semantic relationship of items. A graph is constructed per modality to better capture the information within each one that neighbors together items with similar features on that same modality. An attention mechanism is also employed to extract the most significant features of each modality when merging the information from different modalities for the same item.

User-Item Preference Graph. This graph is constructed based on historical interactions between users and items. The multi-modal information gathered in the previous steps is then propagated through this graph to capture user preferences across modalities.

Model Prediction and Optimization. The final prediction is made based on the inner product of the user and item representations. A loss function is also employed here to optimize the model parameters.

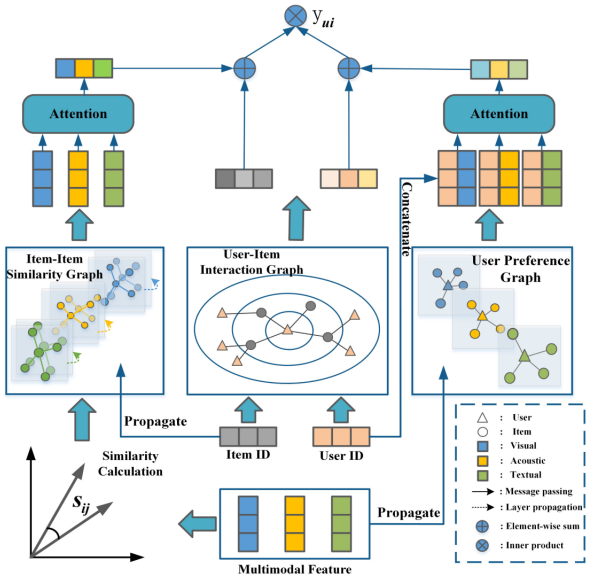

Fig. 8. Architecture of the MMGCN model.

In essence, the presented framework demonstrates the effectiveness of considering multi-modal information regarding the problem of micro-video recommendation.

4 CONCLUSIONS

In conclusion, the recent advancements in micro-video recommendation systems represent significant strides in refining and enhancing existing frameworks. These improvements are primarily focused on developing a more nuanced understanding of user preferences and behaviors, essential for personalizing content in the rapidly evolving domain of micro-video platforms.

At the core of these advancements is the recognition of the complex and multi-faceted nature of user interactions with micro-videos. Traditional recommendation systems, largely based on collaborative filtering and matrix factorization, have evolved to integrate sophisticated techniques like graph convolution networks (GCN) and attention mechanisms. These techniques enable the systems to delve deeper into the rich and multimodal data inherent in micro-videos, encompassing visual, auditory, and textual elements.

The emerging trend is the shift towards a more granular analysis of user preferences. Recent models, such as the MMGCN framework [1], demonstrate an acute awareness of the different modalities within micro-videos and how users interact with them. By constructing item-item modal similarity graphs and user preference graphs, these systems not only enhance item representation but also capture the nuances of user preferences across different modalities. This approach marks a departure from one-size-fits-all recommendation logic, moving towards a more personalized, user-centric model.

Furthermore, the integration of dynamic micro-video trends [2] and fine-grained user interest analysis at clip-level [3] into recommendation systems underscores the industry's drive towards hyper-personalization. By continuously adapting to the shifting trends and evolving user interests, these systems are not just reactive but predictively adaptive, anticipating user needs and preferences.

In essence, the field of micro-video recommendation systems is witnessing a transition from broad-stroke recommendations to intricate, tailored experiences. This shift is powered by advancements in AI and machine learning, which are becoming increasingly adept at interpreting complex user data. As we move forward, the challenge for developers and researchers will be to balance the sophistication of these systems with ethical considerations and user privacy, ensuring that recommendations are not just accurate, but also respectful of user autonomy and consent.

ACKNOWLEDGMENTS

To my cat Lulu, who has nothing to do with this.

REFERENCES

[1] Fei Lei, Zhongqi Cao, Yuning Yang, Yibo Ding, and Cong Zhang. 2023. Learning the User's Deeper Preferences for Multi-Modal Recommendation Systems. *ACM Trans. Multimedia Comput. Commun. Appl.* 19, 3s, Article 138 (feb 2023), 18 pages. https://doi.org/10.1145/3573010

[2] Yujie Lu, Yingxuan Huang, Shengyu Zhang, Wei Han, Hui Chen, Zhou Zhao, and Fei Wu. 2021. Multi-trends Enhanced Dynamic Micro-video Recommendation. *arXiv e-prints*, Article arXiv:2110.03902 (Oct. 2021), arXiv:2110.03902 pages. https://doi.org/10.48550/arXiv.2110.03902 arXiv:2110.03902 [cs.IR]

[3] Yu Shang, Chen Gao, Jiansheng Chen, Depeng Jin, Meng Wang, and Yong Li. 2023. Learning Fine-Grained User Interests for Micro-Video Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (<conf-loc>, <city>Taipei</city>, <country>Taiwan</country>, </conf-loc>) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 433–442. https://doi.org/10.1145/3539618.3591713

[4] Mohd Mustafeez ul Haque and Bonthu Kotaiah. 2023. Hybrid and Classical Models of Recommendation Systems- A Review. In *2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*. 1–6. https://doi.org/10.1109/SCEECS57921.2023.10063125