Enhancing Sequential Recommendation System For MOOCs Based On Heterogeneous Information Networks

Long Nguyen[1,3], Thu Nguyen [2,3], Khoa Tan-Vo[2,3], Tu-Anh Nguyen Hoang[2,3], Tri Nguyen[4], Ngoc-Thanh Dinh[5]
[1] *Faculty of Computer Science, University of Information Technology, Ho Chi Minh City, Vietnam*
[2] *Faculty of Information Science and Engineering, University of Information Technology, Ho Chi Minh City, Vietnam*
[3] *Vietnam National University, Ho Chi Minh City, Vietnam*
[4] *Aalto University, Finland*
[5]*The Industrial University of Ho Chi Minh City, Vietnam*
20520239@gm.uit.edu.vn, {thunta, khoavt, anhnht}@uit.edu.vn
hong-tri.nguyen@aalto.fi, thanh_01030050@iuh.edu.vn

***Abstract*—This study introduces a novel recommendation framework for MOOCs (Massive Open Online Courses)that leverages HINs (Heterogeneous Information Networks) to enhance sequential recommendation models. This is the first attempt to utilize HINs in this context, offering a more comprehensive approach than traditional methods that rely solely on homogeneous data sources. The framework employs HERec, an attentive meta-path based approach to capture contextual information and learn nuanced representations of learners and items. By combining HERec with sequential recommendation models like BERT4Rec and SASRec, the system provides personalized course recommendations based on user preferences, implicit relationships, and contextual factors, addressing the challenges of data diversity and sparsity. Experiments demonstrate a significant increase in accuracy, with the highest improvement reaching 12.69%, providing learners with more relevant and personalized course suggestions.**

***Index Terms*—MOOCs, Recommender Systems, Heterogeneous Information Networks, Personalized Learning**

I. INTRODUCTION

The global impact of MOOCs on education has been well documented in recent literature. For instance, [1] emphasizes the potential of MOOCs in assisting universities to attract more students and establish a global presence by providing access to high-quality educational materials and technology. [2] discusses the potential of the MOOC phenomenon to engage a large and diverse audience with varying learning needs, emphasizing their importance from a lifelong learning perspective. When it comes to MOOC platforms, Coursera, Udacity, and edX are recognized as the most popular, offering the widest range of courses and boasting the highest number of enrolled students [3]. These platforms have played a crucial role in expanding access to education, enabling learners from all over the world to participate in courses across various domains.

Recommender system in MOOC have garnered significant attention from research. A systematic literature review in [4] discusses how recommender systems in MOOCs can help learners find relevant learning objects or elements and inform the design and creation of MOOC. Similarly, research published in [5] highlights that effective and relevant course recommendations can improve learning outcomes, emphasizing the need for recommendation models and algorithms that can be effectively applied to the MOOC platform. In this research, we focus on enhancing the performance of recommender systems. To achieve this, we need to address several challenges:

- **Data diversity**: In MOOC systems, users interact with courses that belong to various categories, and they may watch multiple videos of a course. There are numerous types of entities and relationships in MOOCs, and effectively leveraging this information can help model user preferences, contributing to improved accuracy of recommendation models.
- **Data sparsity**: Despite possessing vast amounts of user and course data, MOOCs face the challenge of data sparsity due to insufficient interaction density between learners and courses [6]. Learners typically engage in only a few courses aligned with their individual needs and interests, resulting in a very sparse user-course interaction matrix. This poses difficulties in building accurate and effective recommendation models.

When it comes to recommendation models, BERT4Rec and SASRec [7], [8] are two prominent studies that perform well with various datasets. However, in the context of MOOCs, they have not yet been able to address the two aforementioned challenges. Specifically, they do not leverage the data diversity from MOOCs, focusing only on the user-course relationship, and are prone to data sparsity as users cannot learn all courses. This leads to BERT4Rec and SASRec not achieving accuracy when applied to MOOC datasets (as shown in Table II).

To address these challenges, this research presents two contributions:

1) We propose leveraging HERec (Heterogeneous Embedding for Recommendation) [9], a powerful approach that

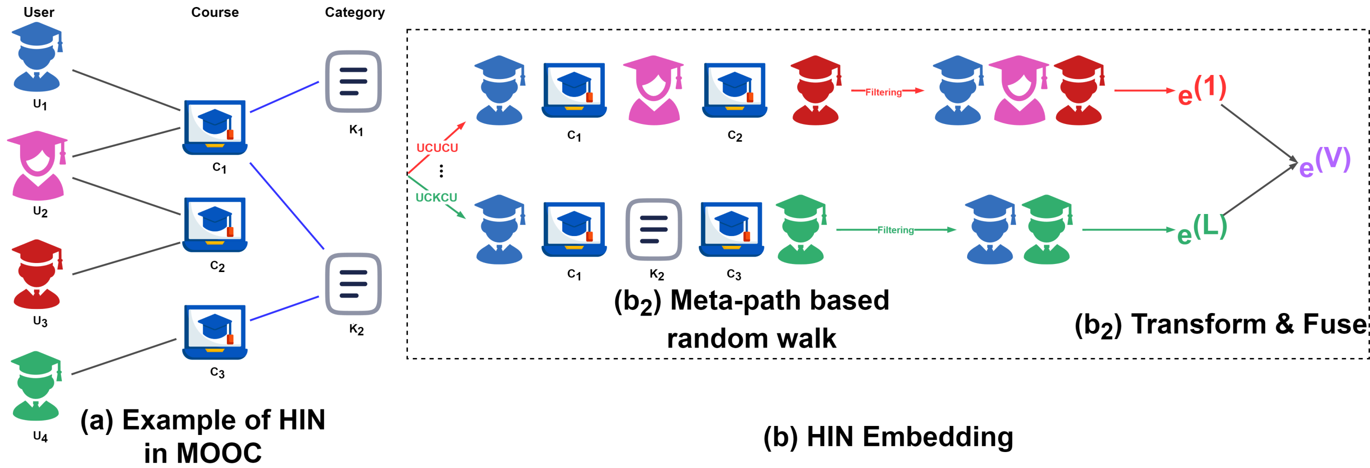

Fig. 1. Workflow Diagram of the HERec Algorithm

excels in handling diverse and sparse data. HERec constructs a knowledge graph to represent various entities (e.g., learners, courses, videos) and their relationships (e.g., ratings, reviews, views). By learning embeddings for these entities and relations, HERec captures the semantic meaning and underlying connections within the MOOC ecosystem. Furthermore, HERec's ability to aggregate information from multiple sources and propagate it through the knowledge graph mitigates the impact of data sparsity.

2) Building upon this foundation, we introduce two novel frameworks, **HERec-SASRec** and **HERec-BERT4Rec**, by integrating HERec with state-of-the-art sequential learning models BERT4Rec and SASRec. This synergy leverages HERec's ability to model complex data while capitalizing on the Transformer architecture of BERT4Rec and SASRec to understand and predict user behavior sequences.

To rigorously evaluate the efficacy of our proposed frameworks, we conduct extensive experiments using real-world data meticulously collected from the XuetangX platform. Through a comprehensive comparative analysis, we assess the impact of incorporating heterogeneous network embeddings on the performance of course recommendation models for MOOC platforms. This evaluation allows us to definitively gauge the effectiveness of our approach in enhancing recommendation accuracy and personalization.

The structure of the remaining sections of this paper is as follows: Section 2 introduces related works. Section 3 provides preliminary knowledge. Subsequently, in Section 4, we present our framework. Section 5 details the experiments and analysis. Finally, conclusions are presented in Section 6.

II. RELATED WORK

A. Sequential Recommendation

Traditional recommender systems, such as collaborative filtering (CF) and content-based filtering (CBF), have achieved

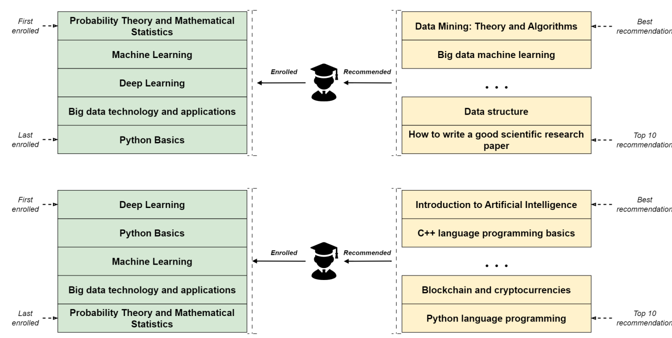

Fig. 2. Impact of course enrollment order on Sequential Recommendation

significant success in various domains [10]. However, when applied to MOOCs, these methods encounter limitations due to the unique characteristics of this learning environment. CF techniques, which rely on user-item interactions, struggle with data sparsity prevalent in MOOCs due to the vast number of courses and dynamic user enrollment patterns. Additionally, CBF approaches, which focus on course content similarity, fail to capture the intricate knowledge structure and prerequisite relationships between courses, potentially recommending options beyond a user's current skill level [11].

To address these shortcomings, recent research has explored the application of sequential recommendation systems for MOOCs. These models excel at capturing the temporal dynamics inherent in user behavior, enabling the system to not only account for a user's current interests but also dynamically adapt to their evolving learning trajectory. To illustrate sequential recommendation, Fig.2 clearly shows that when the order of courses registered in the input sequence changes, the resulting recommended courses for the user also change. Two prominent examples of sequential recommendation models that have shown promise in the context of MOOCs are BERT4Rec and SASRec:

- **BERT4Rec** [7]: This model leverages the power of the Transformer architecture, specifically employing a bidirectional self-attention mechanism. This mechanism allows BERT4Rec to analyze user interactions within a sequential context, not only identifying a user's current course preferences but also understanding the relationships between these courses within their learning path. By considering the order in which users interact with courses, BERT4Rec can capture implicit knowledge dependencies and recommend courses that build upon previously acquired knowledge. Additionally, BERT4Rec benefits from its ability to jointly model various contextual information sources, such as course descriptions, user profiles, and potentially knowledge graph embeddings, leading to a more comprehensive understanding of user intent.

- **SASRec (Self-Attention based Sequential Recommendation)** [8]: This model utilizes a self-attention mechanism specifically designed for sequential recommendation tasks. Unlike BERT4Rec's bidirectional approach, SASRec focuses on capturing the dependencies between courses within a user's interaction sequence in a unidirectional manner. By analyzing these dependencies, SASRec can prioritize courses that logically follow the knowledge acquisition path established through a user's previous course selections. This unidirectional approach allows for efficient model training and inference, making SASRec well-suited for real-world MOOC recommendation systems.

These sequential recommendation models offer significant advancements over traditional methods by incorporating the temporal dimension of user behavior. By understanding the order and relationships between course interactions, BERT4Rec and SASRec can generate more personalized and relevant course recommendations for MOOC learners, ultimately contributing to a more successful learning experience.

B. Heterogeneous Information Networks

While sequential recommendation systems address the temporal dynamics of user behavior, they often lack the ability to capture the rich context within a MOOC ecosystem. This context encompasses not only the courses themselves but also the underlying knowledge structure, prerequisite relationships, learner profiles, and other relevant entities. To address this limitation, recent research has explored the application of HINs for MOOC recommendation.

HINs are a type of network that contains multiple types of nodes and edges. They represent complex relationships in data, capturing the rich semantic meaning of different structural types of objects and links within a network. This multi-typed interconnectedness allows for more nuanced analysis and mining of data, providing a more comprehensive understanding of the underlying structures and relationships [12]. However, applying HINs to real-world recommendation systems presents challenges. One key challenge is data sparsity, which can occur when the number of edges between entities is insufficient to learn robust representations. Additionally, designing effective methods for incorporating diverse data sources and complex relationships within the HIN can be computationally expensive [13] [14] [15].

Heterogeneous Information Network (HIN) [16]: An HIN is denoted as $G = (V, E)$, comprising a set of objects V and a set of links E. An HIN is also associated with an object type mapping function $\phi : V \to A$ and a link type mapping function $\psi : E \to R$. Here, A and R represent predefined sets of object types and link types, where $|A| + |R| > 2$.

Network Schema [17], [18]: A network schema can be represented by the notation $S = (A; R)$. This schema serves as a general reference framework for an information network $G = (V, E)$ with object type mapping functions $\phi : V \to A$ and link type mapping functions $\psi : E \to R$. As a result, we have a directed graph defined over the object types A, with edges representing relationships from the set R.

Example. As shown in Fig.1(a), we have depicted the layout of course recommendation systems on MOOCs using HINs. We further present its corresponding network schema in Fig.1(a), comprising various types of objects. There exist different types of links between objects to represent various relationships. A user-user link indicates the friendship relationship between two users, while a user-course link indicates the enrollment relationship.

Meta-path [19]: A meta-path is a crucial concept in HINs, representing the types of paths between different types of objects.

$$P = A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \cdots \xrightarrow{R_k} A_{k+1}$$

where A_i are object types and R_i are relation types. Meta-paths help create new representation spaces for objects and relationships, thereby improving the effectiveness of machine learning algorithms in HINs.

Common meta-paths between entities in a MOOC HIN can be illustrated as follows: $U \xrightarrow{\text{learn}} C \xrightarrow{\text{learn}^{-1}} U$, indicating that two different users are connected by participating in the same course. Here, learn represents the action of a user U learning a course C, and learn^{-1} is the inverse relationship, representing the course C being learned by the user U. Another example is $U \xrightarrow{\text{watch}} V \xrightarrow{\text{part of}} C \xrightarrow{\text{contains}} K$, showing that users are related through watching videos that are part of the same course.

Recently, HINs have become a mainstream method to model various complex interaction systems [12]. In particular, they have been applied in recommendation systems to depict complex and heterogeneous recommendation settings.

III. PRELIMINARY

A. Personalized Course Recommendation in MOOCs

The description of the problem of building a course recommendation system in MOOCs environments is as follows:

In the realm of MOOCs, there exists the task of constructing an effective course recommendation system. This system comprises the following components:

- Users (U): The set of individuals participating in online courses. $U = \{u_1, u_2, \ldots, u_U\}$

- Courses (C): The collection of courses offered within MOOCs. $C = \{c_1, c_2, \ldots, c_C\}$
- Videos (V): The set of videos contained within each course. $V = \{v_1, v_2, \ldots, v_V\}$
- Categories (K): The categories of knowledge related to the content of videos and courses. $K = \{k_1, k_2, \ldots, k_K\}$

Each course C_i is composed of multiple videos, with each video containing several knowledge concepts. From the user's perspective, an individual U_j will enroll in a course, watch its videos, and interact with some related knowledge concepts to learn the presented content.

The objective of the problem is to enhance the learning experience of users by automatically recommending relevant knowledge concepts. This is based on the users' preferences accumulated from their historical learning experiences.

B. HIN Embedding with HERec

Our first contribution is the utilization of HERec to learn HIN embeddings to address the existing challenges in MOOCs. As shown in Fig.1(b) ,the specific steps involved are as follows:

1) Meta-path based Random Walk:

- **Meta-path Definition:** We identify crucial meta-paths to uncover relationships between courses (Like shown in b2 of Fig.1 focus on users) . For instance:
  - CUC: Course - User - Course (courses taken by similar users)
  - CUVUC: Course - User - Video - User - Course (courses with similar video views by users)
  - CKC: Course - Concept - Course (courses related by concepts)
  - CKCKC: Course - Concept - Course - Concept - Course (courses with deeper conceptual connections)
- **Random Walk Execution:** Starting from a course node, we perform random walks along the defined meta-paths to generate node sequences.

2) Node Filtering: Filter irrelevant nodes to focus on training embeddings for important nodes.

3) Optimization: The optimization goal of HERec to learn HIN embedding methods is to learn the best embeddings for the nodes, maximizing the probability of neighboring nodes within a fixed-length window. The optimization objective function is expressed as follows:

$$\max_f \sum_{u \in V} \log P_r(N_u \mid f(u)),$$

where $f : V \to \mathbb{R}^d$ is the mapping function to be learned to map each node to a d-dimensional feature space, and $N_u \subset V$ is the neighborhood of node u based on a specific meta-path. The embedding learning process is carried out by applying the stochastic gradient descent (SGD) method.

4) Outcome: After transform and fuse, we obtained high-quality course embeddings to form **meta-information** to integrate into the new architecture as shown in the Fig.3.

IV. PROPOSED METHODS

Our study proposes an improvement to the sequential recommendation domain, specifically targeting models like SASRec and BERT4Rec. The proposed architecture is illustrated in Fig. 3. The core modification lies in the embedding layer, where we incorporate three components: item identifier embeddings, positional embeddings, and item meta-information embeddings. Similar to BERT4Rec, our architecture utilizes a Transformer layer preceded by an embedding layer. The standard BERT4Rec embedding layer consists of: (i) an item identifier embedding matrix (denoted by $E_v \in \mathbb{R}^{V \times d}$) and (ii) a positional embedding matrix ($E_p \in \mathbb{R}^{N \times d}$) capturing the position of items within a sequence (where N represents the maximum sequence length). At each step t of the Transformer layer, the input is defined as the sum ($h_t^0 = e_t + p_t$), where e_t (derived from $E_v^{v_t}$) represents the item embedding for item v_t and p_t (derived from $E_p^{(t)}$) represents the positional embedding.

To integrate item meta-information, we deviate from BERT4Rec's approach of using additional meta-information embeddings. Instead, we introduce a novel approach that incorporates meta-information from HERec model Fig.1(b) shown example of HERec items embedding. This involves:

An embedding generated from HERec ($H_v \in \mathbb{R}^{V \times d_1}$), where d_1 represents the embedding dimension size. A transformation matrix ($TM \in \mathbb{R}^{d \times d_1}$) that allows HERec embeddings of any size (d_1) to be seamlessly integrated. The final HERec embedding (h_{et}) for item v_t with a dimension size of d, obtained through $h_{et} = H_v^{v_t} \cdot TM$.

By incorporating h_{et} as an additional summand ($h_t^0 = e_t + p_t + h_{et}$) into the Transformer layer's input at each step, we effectively leverage item meta-information from HERec.

Finally, consistent with BERT4Rec, the final Transformer layer input is masked following the Cloze task for training data preparation, where items for prediction within the sequence are randomly masked.

V. EXPERIMENTS

A. Dataset & Task Setting

Our experiments leverage data from MOOCCUBEX Dataset from the XuetangX[1] MOOC platform in China. This data encompasses user information, their enrolled courses, and the corresponding course categories. Table I presents a comprehensive overview of the MOOCCUBEX dataset, highlighting its key meta-features. The dataset encompasses a substantial user base of 106,388 individuals who have enrolled in a diverse selection of 746 courses. This translates to a total of 683,442 user-course enrollment relationships, indicating a significant level of interaction within the platform. Furthermore, the dataset categorizes these courses into 78 distinct categories, with 944 course-category relationships established, providing a structured representation of the course domain knowledge. Overall, the MOOCCUBEX dataset offers a rich and varied

[1] https://www.xuetangx.com/

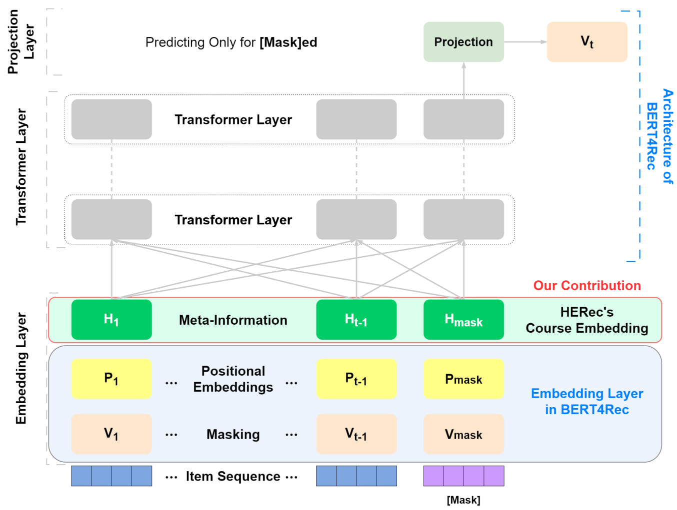

Fig. 3. Framework proposed added HERec model to embedding layer end BERT4Rec for Recommendation System

landscape for exploring user behavior, course preferences, and category-based interactions within the MOOC environment.

We construct training and test datasets from this data. To create training and test instances, we combine a target course with a user's sequence of previously enrolled courses. This sequence reflects the user's learning history. During training, each instance presents a sequence culminating in a target course (to be predicted) preceded by historical courses. During testing, the target course in a user's test set represents a course potentially related to their enrolled courses. The historical course genres in the test set serve as a reference for identifying corresponding courses from the training set belonging to the same user. Additionally, within the test set, we employ negative sampling within the test set. Each positive instance (user-course interaction) is paired with 99 randomly sampled negative instances (courses the user did not interact with), following the same negative sampling strategy used in BERT4Rec [7].

TABLE I
META FEATURES OF DATASET

Node	Count	Description
user	106388	Users enrolled
course	746	Courses have been enrolled
relation user-course	683442	Number of course enrollment
category	78	Category of the course
relation course-category	944	Number of course and category relationships

B. Evaluation Metrics

We assess the quality of the recommendation ranking lists generated by all models using a set of standard evaluation metrics: Hit Ratio (HR) and Normalized Discounted Cumulative Gain (NDCG). Due to the single ground truth item per user in our setting, HR@k is equivalent to Recall@k and proportional to Precision@k. We report HR and NDCG at three different values of k (1, 5, and 10), which represents the number of top-ranked items we consider in the evaluation. For example, HR@5 measures the proportion of users for whom the correct recommendation appears within the top 5 items in the ranking list. Similarly, NDCG@5 evaluates the ranking quality of the top 5 recommendations, taking into account the position of the correct item.

In all these metrics, higher values indicate better model performance, with a higher HR@k suggesting that the model is better at retrieving the correct recommendation, and a higher NDCG@k suggesting that the model is better at ranking the correct recommendation higher in the list. The formulas for HR@k and NDCG@k are as follows:

$$\text{HR@}k = \frac{\sum_{u=1}^{U} \text{hit}_u}{U}$$

$$\text{NDCG@}k = \frac{\text{DCG@}k}{\text{IDCG@}k}$$

where:

$$\text{DCG@}k = \sum_{i=1}^{k} \frac{2^{\text{rel}_i} - 1}{\log_2(i+1)}$$

$$\text{IDCG@}k = \sum_{i=1}^{|\text{REL}|} \frac{2^{\text{rel}_i} - 1}{\log_2(i+1)}$$

and:

- $\text{hit}_u = 1$ if the correct item is in the top k recommendations for user u, otherwise $\text{hit}_u = 0$
- U is the total number of users
- rel_i is the relevance of the item at position i (in our case, 1 if it is the correct item, 0 otherwise)
- $|\text{REL}|$ is the number of relevant items (in our case, 1)

C. Baseline Models Comparison

When employing HERec for input data embedding, our goal is to generate output vectors representing course embeddings. To achieve this, we utilize various meta-paths: **CUC** captures course similarity based on shared users (e.g., "Data Science Fundamentals" and "Applied Machine Learning" are connected if the same users enroll in both); **CKC** captures similarity based on shared knowledge concepts; **CUCUC** and **CKCKC** extend these relationships by considering a wider range of users, courses, and concepts. After transforming the meta-path information as illustrated in Fig.1(b), we integrate it into the embedding layer. This process is performed across various hidden dimensions (d = 16, 32, 64, 128, 256), which correspond to the dimensions used in the BERT4Rec and SASRec models. By incorporating information from these diverse meta-paths, our model gains a more comprehensive understanding of course relationships, leading to richer course embeddings.

TABLE II
COMPARISON OF RECOMMENDATION METRICS
FOR OUR PROPOSED METHOD VS. BERT4REC AND SASREC

Metric	SAS			BERT		
	SASRec	HERec-SASRec	Improve	BERT4Rec	HERec-BERT4Rec	Improve
HR@1	0,2028	0,2182	**+7,59%**	0,2062	0,2275	**+10,33%**
HR@5	0,5004	0,5194	**+3,80%**	0,5145	0,5404	**+5,03%**
HR@10	0,6604	0,6854	**+3,79%**	0,6775	0,7031	**+3,78%**
NDCG@5	0,3557	0,3757	**+5,62%**	0,3656	0,412	**+12,69%**
NDCG@10	0,4076	0,4376	**+7,36%**	0,4184	0,4413	**+5,47%**

D. Result

Table II shows the best result of our experiment. The result of this study demonstrate significant performance improvements in combined models utilizing BERT (BERT4Rec and HERec-BERT4Rec) compared to individual models like BPR and SASRec. Across all top-N metrics including HR@1, HR@5, HR@10, NDCG@5, and NDCG@10, the combined models outperform consistently. Our proposed frameworks, **HERec-SASRec** and **HERec-BERT4Rec**, demonstrate effective performance enhancements compared to their original versions at optimal configurations. The highest observed improvement is 12.69%, with the lowest being 3.79%, achieved without substantial increases in training time.

VI. CONCLUSION

Our study integrates BERT4Rec and SASRec models within a HERec framework to improve MOOC recommendation accuracy. Empirical results demonstrate the positive impact of HINs on sequential recommendation models by capturing diverse information and enriching user-item interaction context. This integration of powerful sequential models like BERT4Rec and SASRec with HINs represents a promising direction for future research in recommendation systems, especially for complex domains like MOOCs.

ACKNOWLEDGMENT

This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under grant number DS2024-26-02

REFERENCES

[1] A. M. Mutawa, "Perspective chapter: Moocs at higher education: current state and future trends," 2023.

[2] M. Buhl and L. B. Andreasen, "Learning potentials and educational challenges of massive open online courses (moocs) in lifelong learning," pp. 151–160, 2018.

[3] S. Taneja and A. Goel, "Mooc providers and their strategies," *International Journal of Computer Science and Mobile Computing*, vol. 3, no. 5, pp. 222–228, 2014.

[4] K. Najmani, N. Sael, A. Zellou *et al.*, "A systematic literature review on recommender systems for moocs," *Ingenierie des Systemes d'Information*, vol. 27, no. 6, p. 895, 2022.

[5] H. Zhang, T. Huang, Z. Lv, S. Liu, and Z. Zhou, "Mcrs: A course recommendation system for moocs," *Multimedia Tools and Applications*, vol. 77, pp. 7051–7069, 2018.

[6] Y. Zhang, Y. Gao, Y. Zhou, J. Wang, and M. Yin, "Mrmlrec: A two-stage approach for addressing data sparsity in mooc video recommendation (student abstract)," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 21, 2024, pp. 23 709–23 711.

[7] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang, "Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer," in *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019, pp. 1441–1450.

[8] W.-C. Kang and J. McAuley, "Self-attentive sequential recommendation," in *2018 IEEE international conference on data mining (ICDM)*. IEEE, 2018, pp. 197–206.

[9] C. Shi, B. Hu, W. X. Zhao, and S. Y. Philip, "Heterogeneous information network embedding for recommendation," *IEEE transactions on knowledge and data engineering*, vol. 31, no. 2, pp. 357–370, 2018.

[10] D. Roy and M. Dutta, "A systematic review and research perspective on recommender systems," *Journal of Big Data*, vol. 9, no. 1, p. 59, 2022.

[11] P. Castells and D. Jannach, "Recommender systems: A primer," *arXiv preprint arXiv:2302.02579*, 2023.

[12] C. Shi, Y. Li, J. Zhang, Y. Sun, and S. Y. Philip, "A survey of heterogeneous information network analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 1, pp. 17–37, 2016.

[13] B. Yan, Y. Cao, H. Wang, W. Yang, J. Du, and C. Shi, "Federated heterogeneous graph neural network for privacy-preserving recommendation," *arXiv preprint arXiv:2310.11730*, 2023.

[14] J. Zhang, S. Wang, Z. Chen, X. Yang, and Z. Peng, "Dependency relationships-enhanced attentive group recommendation in hins," *arXiv preprint arXiv:2311.11239*, 2023.

[15] Y. Yang, Z. Guan, J. Li, W. Zhao, J. Cui, and Q. Wang, "Interpretable and efficient heterogeneous graph convolutional network," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 2, pp. 1637–1650, 2021.

[16] Y. Sun and J. Han, *Mining heterogeneous information networks: principles and methodologies*. Morgan & Claypool Publishers, 2012.

[17] ——, "Mining heterogeneous information networks: a structural analysis approach," *ACM SIGKDD explorations newsletter*, vol. 14, no. 2, pp. 20–28, 2013.

[18] Y. Sun, Y. Yu, and J. Han, "Ranking-based clustering of heterogeneous information networks with star network schema," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 797–806.

[19] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "Pathsim: Meta path-based top-k similarity search in heterogeneous information networks," *Proceedings of the VLDB Endowment*, vol. 4, no. 11, pp. 992–1003, 2011.