Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.049129

ARTICLE

LKPNR: Large Language Models and Knowledge Graph for Personalized News Recommendation Framework

Hao Chen[#], Runfeng Xie[#], Xiangyang Cui, Zhou Yan, Xin Wang, Zhanwei Xuan[*] and Kai Zhang[*]

State Key Laboratory of Communication Content Cognition, People's Daily Online, Beijing, 100733, China

*Corresponding Authors: Zhanwei Xuan. Email: xuanzhanwei@people.cn; Kai Zhang. Email: zhangkai@people.cn

[#]These authors contributed equally

Received: 28 December 2023 Accepted: 15 April 2024 Published: 20 June 2024

ABSTRACT

Accurately recommending candidate news to users is a basic challenge of personalized news recommendation systems. Traditional methods are usually difficult to learn and acquire complex semantic information in news texts, resulting in unsatisfactory recommendation results. Besides, these traditional methods are more friendly to active users with rich historical behaviors. However, they can not effectively solve the long tail problem of inactive users. To address these issues, this research presents a novel general framework that combines Large Language Models (LLM) and Knowledge Graphs (KG) into traditional methods. To learn the contextual information of news text, we use LLMs' powerful text understanding ability to generate news representations with rich semantic information, and then, the generated news representations are used to enhance the news encoding in traditional methods. In addition, multi-hops relationship of news entities is mined and the structural information of news is encoded using KG, thus alleviating the challenge of long-tail distribution. Experimental results demonstrate that compared with various traditional models, on evaluation indicators such as AUC, MRR, nDCG@5 and nDCG@10, the framework significantly improves the recommendation performance. The successful integration of LLM and KG in our framework has established a feasible way for achieving more accurate personalized news recommendation. Our code is available at https://github.com/Xuan-ZW/LKPNR.

KEYWORDS

Large language models; news recommendation; knowledge graphs (KG)

1 Introduction

With the exponential growth of the Internet, an increasing number of individuals are opting to access the most current global news via online platforms, including MSN News. However, users often find themselves overwhelmed by the sheer volume of available news content. Consequently, the imperative for an effective news recommendation system becomes evident, as it serves as a crucial tool in assisting users to navigate through this vast information landscape. Such a system not only aids users in filtering copious amounts of news but also employs personalized algorithms to proactively present news items aligning with users' genuine interests, thereby significantly enhancing the fulfillment of their informational requirements [1].

The current research in this field primarily emphasizes the perspective of representation learning [2], aiming to enhance the learning of user and news representations separately. To illustrate this, we take the MIND dataset [3] as a case study. The user's behavioral data within this dataset comprises a sequence of historical clicks on news items, each of which is composed of a title, abstract, category and other related information. Earlier researches, exemplified by NAML [4] and NRMS [5], have been employed to achieve improved news representations. The methods often leverage convolutional neural network (CNN) [6] and long short-term memory (LSTM) [7] for feature extraction, coupled with feature fusion using attention mechanisms. However, these existing studies exhibit certain limitations that deserve attention. One such problem is insufficient news text feature extraction, this deficiency hampers the effectiveness of the news representation process. Neglect of news popularity and interconnections is another challenge, this neglect can potentially lead the news recommender system into the challenging scenario of the long-tail problem.

With the advent of ChatGPT, the realm of natural language processing has seamlessly ushered in the era of Large Language models (LLMs). Notably, significant models like ChatGLM2 [8], LLAMA2 [9], and RWKV [10] have surfaced within the open-source community. LLMs, having undergone pre-training on vast corpora of textual data, exhibit the ability to swiftly acclimate to the data distribution pertaining to downstream tasks. Leveraging the exceptional language modeling proficiency of LLMs, they adeptly uncover intricate linguistic relationships and semantic nuances inherent within the text. This capacity allows for a more robust contextual integration, thereby augmenting text comprehension and facilitating the extraction of information-rich semantic features.

The long-tail problem [11] in recommendation systems is that a significant majority, approximately 80%, of user clicks are concentrated on a mere 20% of popular items. Consequently, this tendency results in recommendation systems favoring these popular items, often overlooking less popular ones, which, over time, detrimentally impacts the overall effectiveness of recommendations. To address this long-tail challenge, recent research [12,13] has explored the incorporation of knowledge graphs (KGs) as supplementary information for recommender systems. This innovative approach leverages graph-based learning [14] to establish meaningful relationships among diverse items within the system. Subsequently, this method harnesses additional item-specific information to enhance the representation of long-tailed items, effectively mitigating the issue of inadequate representation learning for such items.

To address the aforementioned challenges, we introduce LKPNR (LLM and KG for Personalized News Recommendation framework), a personalized news recommendation framework that links general news recommendation models with the integration of KGs and LLMs. Capitalizing on the robust text comprehension abilities of the LLM, we generate news representations imbued with comprehensive semantic information for each news item, thereby mitigating the shortcomings associated with the limited feature extraction capabilities inherent in general news recommendation models. Concurrently, the incorporation of subgraph structural representations, mined through multi-hop inference within KG, serves to alleviate the issue of long-tail distribution prevalent in news recommendation. By harnessing the strengths of both LLM and KG, we observe a substantial enhancement in the model's performance. The differences between our proposed framework and the general news recommendation model are shown in Fig. 1. To summarize, our contributions are listed as follows:

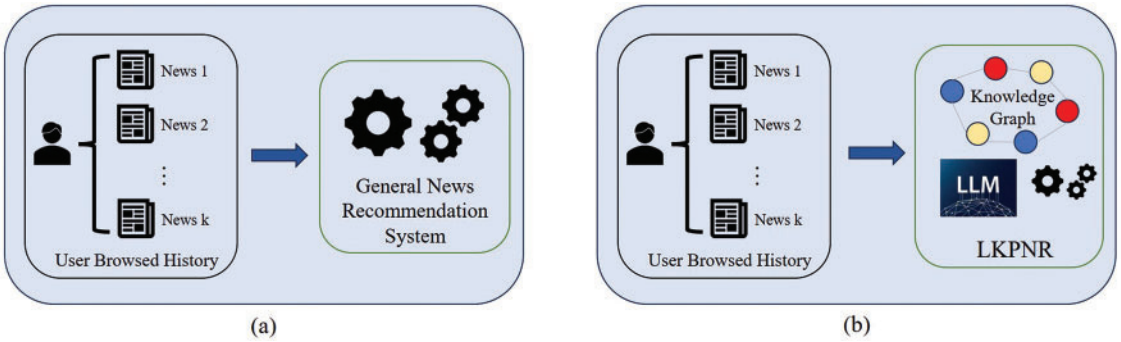


Figure 1: Differences between traditional news recommendation model (a) and LKPNR framework (b)

- We propose LKPNR, a personalized news recommendation framework that fuses general news recommendation models with KG and LLM. To the best of our knowledge, this is the first work that combines both KG and LLM in the news recommendation domain.
- LKPNR can be flexibly combined with various general news recommendation models. Leveraging LLM`s powerful text understanding and the news graph structural relationships contained in the KG to inject additional information into general news recommendation models.
- Experiments on the MIND dataset show that LKPNR can significantly improve the performance of general news recommendation models.

2 Related Work

2.1 *General News Recommendation Model*

General news recommendation models typically involve encoding various aspects of news, such as title and abstract independently. These encoded representations are then interacted with separately to create a comprehensive news representation. In a similar manner, historical news browsing sequences are encoded and integrated to form a user representation. Subsequently, the similarity between this user representation and the representations of candidate news items is computed, enabling the prediction of whether the user would find these candidates interested or not. Okura et al. [15] extracted news features through a denoising self-encoder, while user sequences are derived using RNN to obtain user features. Lian et al. proposed DFM [16] using multi-channel inception blocks and attention mechanism to tackle the issue of data diversity. An et al. [17] employed GRU to model both long-term and short-term interests, resulting in improved user representations. Vaswani et al. extended the attention mechanism [18] paradigm in several ways [4,19]. Wang et al. [20] employed fine-grained interest matching using dilated convolution and 3D convolution techniques.

Traditional models usually encode part of news information (such as headlines and abstracts) independently, which may lead to the underutilization of their internal relations and interactions. In addition, although the user's historical browsing sequences is encoded and integrated, these methods may not fully consider the dynamic interest changes of users, resulting in limited accuracy of recommendation results.

2.2 *LLM-Powered News Recommendation Model*

With the remarkable performance exhibited by LLM across diverse domains, researchers have embarked on an exploration of its potential within the recommendation domain. Kang et al. [21]

conducted a comparative study encompassing traditional collaborative filtering methods and LLM for user rating prediction, examining zero-shot, few-shot, and fine-tuned scenarios. Their research revealed intriguing insights. Likewise, Hou et al. [22] devised various prompts to address the ranking predicament in recommender systems. Gao et al. [23] transformed user profiles and historical interactions into prompts, leveraging ChatGPT's in-context learning capabilities for recommendation.

The above methods utilize LLM's in-context learning capabilities by constructing prompts to cope with downstream tasks in the general recommendation domain, but the performance is far inferior to that of traditional ID embedding-based fine-tuning methods. In the news recommendation domain, LLM is also beginning to combine with traditional models. Liu et al. [24] used ChatGPT to generate user profiles and augment the news text, combining with a traditional recommendation model to achieve better results. Li et al. [25], based on the traditional recommendation model, generated news representations by using LLM as a news encoder directly to complete the news recommendation.

2.3 KG-Powered News Recommendation Model

Recommendation systems predominantly operate within a graph structure due to the information-rich nature of the data, which leads to the widespread adoption of GNN within the realm of news recommendation. Wang et al. [26] proposed DKN, which constructs a graph of entities within the input news and was utilized to seek neighboring nodes for expanding news information. Another significant contribution by Mao et al. [27] involves the integration of a GCN with the user's browsing history and leverages both LSTM and attention mechanisms to extract textual information. Yang et al. [28] employed a gated GNN, combining global representations from other users with local representations to enhance the personalized recommender system. Furthermore, several research endeavors aim to amplify representation and inference capabilities by linking LLM and KG. Sun et al. [29] utilized attention mechanism to facilitate interaction between all tokens of the input text and KG entities, thus augmenting the output of the language model with domain-specific knowledge from the KG.

It is evident that the KG within the news domain holds substantial value in terms of structured information and domain-specific knowledge. Its inherent incompleteness and limited language understanding are effectively supplemented by the extensive general knowledge encapsulated within LLM. Notably, there is currently a dearth of work that combines LLM and KG within the news recommendation field. Therefore, we propose LKPNR as a solution to bridge this gap. A significant innovation of LKPNR lies in the clever integration of large language models and knowledge graphs into recommendation systems. It not only makes use of the powerful language understanding ability of large language models, but also integrates structured information in knowledge graphs and specific domain knowledge into recommendation systems.

3 Problem Formulation

The personalized news recommendation system proposed takes the user's news clicks and the KG of the entities within the news as input. The user's news clicks refer to a number of pairs of user and news, denoted as $\mathcal{C} = (\mathfrak{u}, \mathfrak{n}) \subseteq \mathcal{U} \times \mathcal{N}$, where $\mathcal{U}$ is the user set and $\mathcal{N}$ is the news set. The KG of the entities contained in the news set $\mathcal{N}$ is a triple $\mathcal{G} = (\mathrm{h}, \mathrm{r}, \mathrm{t}) \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$, where $\mathrm{h}, \mathrm{t} \in \mathcal{E}$ represent the head and tail entities, and $\mathrm{r} \in \mathcal{R}$ represents the relation between them. Given the training click set $\mathcal{H}$ and test click set $\mathcal{T}$, where $\mathcal{H} \cup \mathcal{T} = \mathcal{C}$ and $\mathcal{H} \cap \mathcal{T} = \phi$. Then the recommendation task can be formulated as learning a matching function $\mathcal{F}$ from the training set $\mathcal{H}$ and predicting the degree of user-news matching in the test set $\mathcal{T}$ to obtain the score $\mathcal{M}$

4 Framework

The overall recommendation system framework consists of three components: The LLM and KG Augmented News Encoder (LK-Aug News Encoder), the LLM and KG Augmented User Encoder (LK-Aug User Encoder), and the Click Predictor. The overall recommendation framework is illustrated in Fig. 2. The lower left corner of figure is the input news data into LKPNR, where the words marked in yellow indicates the entities. Candidate news is encoded by LK-Aug news encoder that combines LLM and KG with traditional general news encoder to obtain news representation. To obtain user representation, LK-Aug User Encoder contains several LK-Aug News Encoders, which encodes user's historical click behaviors.

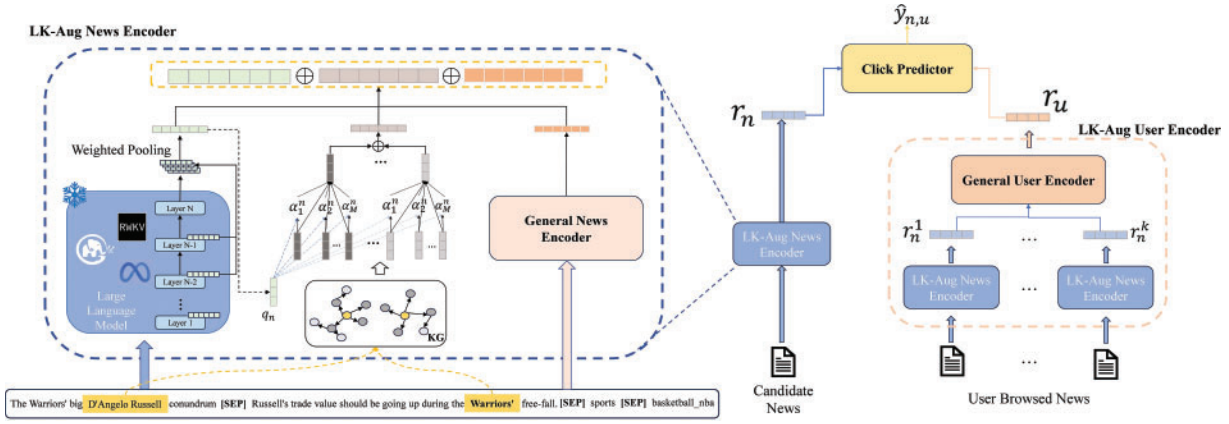


Figure 2: The framework of LKPNR

4.1 LK-Aug News Encoder

The LK-Aug News Encoder is composed of three sub-modules: General News Encoder, LLM-Augmented Encoder and KG-Augmented Encoder.

4.1.1 General News Encoder

The general news encoder is designed to learn a semantic representation of news at the word level, which is achieved through a structure that typically consists of two layers: The word embedding layer and the word fusion layer. The word embedding layer utilizes an embedding matrix $E \in R^{(|v| \times d)}$ to convert each word w occurring in the news title and abstract into an embedding vector e. Here, v denotes the number of words and d denotes the dimension of the word embedding. The fusion layer is a well-designed component in the baseline experiments that interacts with the embedding vectors of individual words via operations such as linear mapping, weighted summing, and vector concatenation. These operations fuse the vectors to produce a generic representation vector r_{GNE} for the given news.

4.1.2 LLM-Aug Encoder

We leverage powerful contextual understanding capability of LLM and rich open-world knowledge to build news representations with comprehensive semantic information for the input news to solve the problem of limited text feature extraction capability in general news recommendation. The input news text is created by concatenating news titles, news abstract, category and sub-category through [SEP], as shown in Fig. 2.

$$S_{HS} = LLM\left(t\right) \tag{1}$$

$$S_{HS_p} = mean\_pooling(S_{HS}) \tag{2}$$

where $LLM(\cdot)$ is the LLM decoder which returns the hidden states of the last four decoder layers, denoted S_{HS}, and S_{HS_p} denotes the output after taking mean pooling strategy on the sequence length dimension of each layer.

After that S_{HS_p} will be weighted and summed to get S_w, this process is shown in Eqs. (1)–(3).

$$S_W = \sum_{i=1}^{4}\left(a_i S^i_{HS_p}\right) \tag{3}$$

where a_i denotes the learnable weights, $S^i_{HS_p}$ denotes the hidden states of the i_{th} layer.

Finally the weighted hidden state S_w will be projected to the text representation space by nonlinear mapping.

$$r_{LLM} = \sigma\left(f_l\left(S_W\right)\right) \tag{4}$$

where σ denotes the activation function, f_l denotes the linear transformation, and r_{LLM} is the enhanced news representation.

4.1.3 KG-Aug Encoder

We feed the enhanced news representation r_{LLM} into the nonlinear layer, which is used to transform r_{LLM} from the textual representation space to the entity representation space.

$$q = \sigma\left(f_s\left(r_{LLM}\right)\right) \tag{5}$$

where f_s is a linear transformation. The query q will have extensive interactions with the entities in the KG.

Generally, the title and abstract of a piece of news will contain several source entities. Considering the 1, 2, $\ldots$, n hop adjacent entities of these source entities, we can extract a subgraph $g^n = (V, R)$ from the external KG, where V is the set of entities in the subgraph, R is the set of edges connecting entities. Taking the k_{th} hop adjacent entity set as an example, $V^k = \left\{v^k_i\right\}_{i=0}^{|M|}$ represents M entities with k hops from the source entity. Then we get the embedding vector corresponding to each entity through the wiki KG, denoted as $X^k = \left\{x^k_i\right\}_{i=0}^{|M|}$. Given a query q and a neighboring entity set X^k with hop count k, the KG-Augmented Encoder interacts the query with each vector x^k_i in X^k to generate a news attention score for the entity, denoted as α^k_i in Eq. (6).

$$\alpha^k_i = Softmax\left(W^T_k\left[q; x^k_i; q \circ x^k_i\right]\right) \tag{6}$$

where $W^T_{\alpha k} \in R^{3l\times 1}$ is a learnable parameter matrix, $\circ$ is the element multiplication, and [;] denotes the vector connection.

The weighted representation $\hat{x}^k$ of a k-hop entity can be computed as:

$$\hat{x}^k = \sum_{i=1}^{M}\alpha^k_i x^k_i \tag{7}$$

After that, we concatenate the weighted representation vectors of each hop and project the concatenated vectors to the news representation space, denoted as the KG representation r_{KG} in Eq. (8).

$$r_{KG} = Q^T\left[\hat{x}^1; \ldots; \hat{x}^n\right] \tag{8}$$

where $Q^T \in R^{nl\times o}$, nl denotes the dimension after the representation vector connection of each hop and o represents the dimension of the projected KG representation.

The final news representation vector r_n is obtained by connecting three news representations above, which is shown as Eq. (9).

$$r_n = [r_{GNE}; r_{KG}; r_{LLM}] \tag{9}$$

4.2 LK-Aug User Encoder

The LK-Aug User Encoder learns representations of users based on their click history on news. This module includes a News Embedding Layer (NEL) and a Representation Fusion Layer (RFL). The NEL obtains the representation of the news browsed by the user through the LK-Aug News Encoder, denoted as $[r_n^1, r_n^2, \ldots, r_n^z]$, where z represents the length of the historical browsing news sequence. The RFL transforms news representation sequences into user representations r_u which is shown as Eqs. (10), (11) through a series of fusion methods, such as concatenation, mapping, attention, etc.

$$\left[r_n^1, r_n^2, \ldots, r_n^k\right] = NEL\left(h_1, h_2, \ldots, h_k\right) \tag{10}$$

$$r_u = RFL\left(\left[r^1, r^2, \ldots, r^k\right]\right) \tag{11}$$

4.3 Click Predictor and Model Training

Given candidate news and user representations r_n and r_u, this module is used to get the matching score between user u and the candidate news n. We compute the dot-product $\hat{y}_{n,u}$ of r_n and r_u as the unnormalized matching scores of users and news in Eq. (12).

$$\hat{y}_{n,u} = \langle r_n, r_u\rangle \tag{12}$$

Following previous general work, we use a negative sampling strategy for model training. For the i_{th} news exposure, we compute its unnormalized matching score as $\hat{y}_{\mathrm{i}}^{+}$. In addition, randomly select K pieces of news as the news that the user does not click, and its unnormalized matching score is $[\hat{y}_{i,1}^{-}, \hat{y}_{i,2}^{-}, \ldots, \hat{y}_{i,K}^{-}]$. We employ Eq. (13) to calculate the normalized matching score.

$$p_i = \frac{exp\left(\hat{y}_i^{+}\right)}{exp\left(\hat{y}_i^{+}\right) + \sum_{j=1}^{K} \exp\left(\hat{y}_{i,j}^{-}\right)} \tag{13}$$

In this way, the click prediction problem can be formulated as a K + 1 classification problem. For the training set $\mathcal{H}$, the loss function $\mathcal{L}$ of model training is the negative log likelihood of all positive samples, which can be described as follows:

$$\mathcal{L} = \sum_{i\in\mathcal{H}} log\left(p_i\right) \tag{14}$$

5 Experiments

5.1 DataSet

We conduct experiments on the MIND dataset, which is a dataset constructed based on user click logs on the MSN online news site. We use the same processing method as Mao et al. [27]. We randomly sample 200 K user click logs from the training and validation sets of the MIND dataset. Given the absence of labels for the test set, we partition the original validation set into two distinct segments: The experimental validation set and the experimental test set. The specific information of the constructed sampled dataset is shown in Table 1.

Table 1: Statistics of MIND-200K dataset

#users	20000	#user in train set	189580
#news	78602	#news in train set	75963
#training logs	595186	#training samples	905297

5.2 Implementation Details and Metrics

We select NAML, NRMS, LSTUR and NPA as our baseline models, setting the maximum sequence length for user browsing history to 50 and adopting a negative sampling rate of $k = 4$. We maintain the parameter configurations of the General News Encoder identical to the baseline. In the LLM-augmented encoder, the hidden states of the LLM are projected to 500 dimensions. For the KG-augmented encoder, the dimension of entity embedding is set to 100. Furthermore, we limit the maximum number of neighboring nodes to 20 per source node, and the traversal depth is constrained to a maximum of 2 hops. Throughout the training process, learning rate employs 1e–4, batch size is set to 64, and early-stop strategy is implemented. All experiments are conducted on the NVIDIA TESLA V100.

To evaluate the model's performance, we use four widely recognized evaluation metrics, specifically, the Area Under the ROC Curve (AUC), Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (nDCG@5 and nDCG@10).

In addition, We also investigated the latest and most influential relevant models, and the PPSR model proposed by Ding et al. is more representative [30]. Therefore, we applied our LKPNR framework to the PPSR model.

5.3 Performance Comparison

Five basic models, NRMS, NAML, LSTUR, NPA and PPSR, were used for comparison. Here, the LLM is Chatglm2-6B. In addition, to further validate the efficacy of our framework design, we conduct ablation experiments, the experimental results are summarized in Table 2. Where Orig. denotes general news recommendation, LKPNR denotes the framework that we proposed, LKPNR (w/o KG) denotes the framework that remove the KG-Augmented Encoder, LKPNR (w/o LLM) denotes the framework that remove the LLM Augmented Encoder.

Table 2: Performance of comparison results

Methods		AUC	MRR	nDCG@5	nDCG@10
NRMS	Orig.	0.6802	0.3316	0.3661	0.4306
	LKPNR	**0.7049**	**0.3492**	**0.3886**	**0.4514**
	LKPNR (w/o KG)	0.6997	0.3441	0.3846	0.4453
	LKPNR (w/o LLM)	0.6845	0.3308	0.3680	0.4310
NAML	Orig.	0.6842	0.3257	0.3621	0.4257
	LKPNR	**0.7023**	**0.3423**	**0.3816**	**0.4440**
	LKPNR (w/o KG)	0.6995	0.3411	0.3810	0.4426
	LKPNR (w/o LLM)	0.6841	0.3319	0.3699	0.4325

(Continued)

Table 2 (continued)

Methods		AUC	MRR	nDCG@5	nDCG@10
LSTUR	Orig.	0.6827	0.3293	0.3643	0.4277
	LKPNR	**0.6980**	**0.3432**	**0.3814**	**0.4437**
	LKPNR (w/o KG)	0.6935	0.3378	0.3749	0.4375
	LKPNR (w/o LLM)	0.6806	0.3278	0.3648	0.4278
NPA	Orig.	0.6757	0.3246	0.3591	0.4219
	LKPNR	**0.6927**	**0.3369**	**0.3764**	**0.4380**
	LKPNR (w/o KG)	0.6918	0.3373	0.3754	0.4374
	LKPNR (w/o LLM)	0.6786	0.3286	0.3628	0.4252
PPSR	Orig.	0.6944	0.3445	0.3834	0.4427
	LKPNR	**0.6955**	**0.3447**	**0.3837**	**0.4439**
	LKPNR (w/o KG)	0.6928	0.3443	0.3835	0.4415
	LKPNR (w/o LLM)	0.6902	0.3426	0.3332	0.4428

The experimental results indicate that LKPNR can improve multiple indicators of models, such as AUC, MRR, nDCG@5, nDCG@10, etc. This performance improvement derives from the fact that the news encoder of the baseline gains better performance through the enhanced semantic information of the LLM and the collaborative information of the KG. As depicted in Table 2, discernible decrements in performance are across the spectrum of ablation variants compared to our complete model, which demonstrates the efficacy of the different components of our model. The removal of the LLM exhibits a substantial impact on the overall performance of the model, demonstrating the effectiveness of the augmenting semantic information for news representation.

5.4 Performance of Different LLM

The characteristics of LLM can vary due to inconsistencies in the proportion of data categories within the training data and model structural design. Consequently, these differences lead to varying capabilities among LLMs, reflected in their open-world knowledge, performance on diverse tasks and so on, which leads to their distinct understanding of text. For instance, ChatGLM2 [8] is trained on the same amount of Chinese and English corpus, and can handle Chinese and English tasks with high quality. LLAMA2 [9], trained on an extensive, high-quality English dataset, demonstrates adeptness in handling various English tasks. RWKV [10] exhibits a quicker reasoning speed and lower computational complexity. In order to explore the impact of various LLMs on news recommendation, we employ three cutting-edge models: ChatGLM2, LLAMA2, and RWKV to generate enhanced news representations on the basis of the baseline model NRMS. Table 3 shows the performance comparison of the enhanced news representation of different LLMs in the recommendation task.

The outcomes of our experiments utilizing these three diverse LLMs are detailed in the table below, and the results indicate that ChatGLM2 provides the most effective enhancement for news recommendation when compared to both LLAMA2 and RWKV. The reason for this phenomenon is that the training data of ChatGLM2 may contain a certain proportion of English news.

Table 3: The performance comparison of the enhanced news representation of different LLMs

Methods		AUC	MRR	nDCG@5	nDCG@10
NRMS	Orig.	0.6802	0.3316	0.3661	0.4306
	LKPNR (ChatGLM2-6B)	**0.7049**	**0.3492**	**0.3886**	**0.4514**
	LKPNR (LlaMA2-13B)	0.6845	0.3370	0.3657	0.4307
	LKPNR (RWKV-7B)	0.6771	0.3300	0.3631	0.4218

5.5 Effectiveness of KG Entity's Query

The incorporation of neighboring entity vectors in the news coding process can be perceived as a mechanism that augments the collaborative information. By fusing vectors of neighboring entities, the KG-Augmented Encoder is able to gather information from multiple related entities and synergistically integrate them into a single encoded representation. The crux of enhancing news representation through the use of KG lies in the efficient extraction of information from all neighboring entities. In KG-Augmented Encoder, the query is obtained by projecting the textual representation of news into the entity representation space.

However, direct weighted summation of all entities at each hop may lead to relatively large information loss. We use multi-head query to consider different representation spaces, i.e., the multiple weighted summation of all entities with different weights can capture collaborative information from different perspectives of the news and entities, and thus improves the representational capability of KG-Augmented Encoder. Based on NRMS, NAML, we implemented the experiments to compare the retrieval performance of the query converted by general news encoder and query converted by LLM. The detailed experimental results are shown in Fig. 3. Where (a) is based on NRMS, (b) is based on NAML. General_Query denotes General News Encoder mapping query, LLM_Query denotes LLM mapping query.

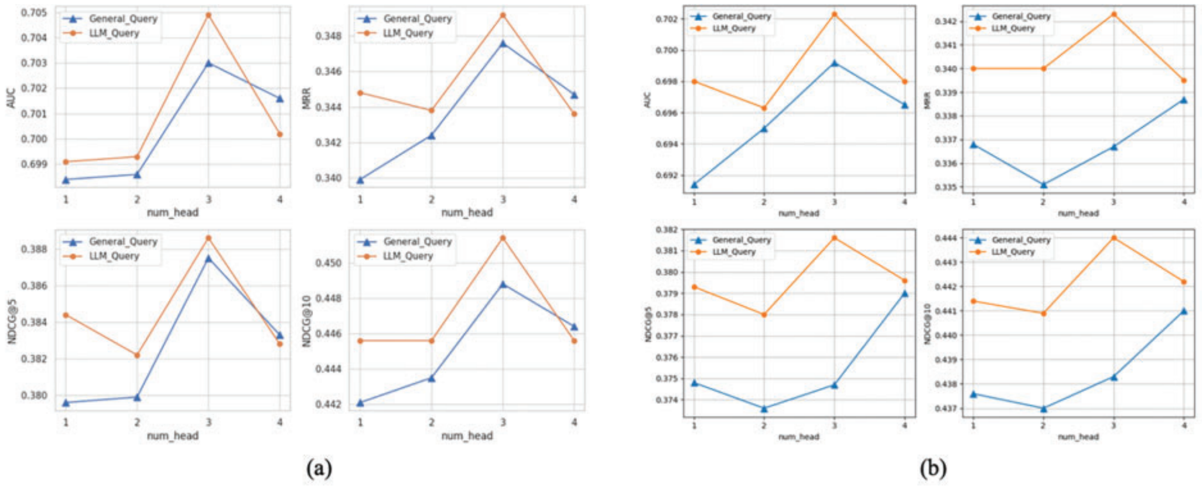


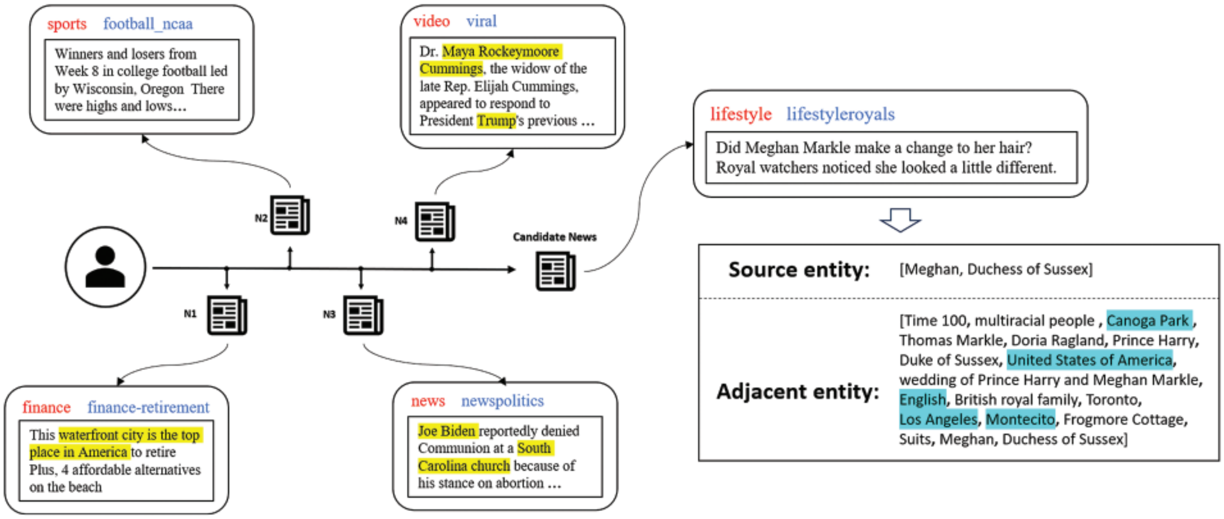
Figure 3: (a) NRMS (b) NAML

The experimental results show that the retrieval performance of LLM mapping query outperforms the General News Encoder mapping query. Compared with the news representation produced by the

General News Encoder, which is limited to the specifics of the news text, the news representation of the LLMs contains some open world knowledge about the news, and is thus able to understand the information of neighboring entities more effectively. In addition, when num_head = 3, the LLM demonstrates its highest proficiency in mapping queries to extract information. It suggests that expanding the representation space of the query vector enhances its capability to gather entity information.

6 Case Study

In this section, we demonstrate the characteristics of LLM and KG Augmented News Encoder using visualization. Fig. 4 shows a user's click order of historical news and current news candidates, as well as the detailed categories, subcategories, titles and abstracts of this news. Before the integration of KG, the user's expression was transformed from the text and category characteristics of these historical news clicks. This candidate news has little correlation with any historical news clicked by users, and the matching degree between the candidate news and the user vector is also very low. After KG integration, give full consideration to adjacent nodes. Fig. 4 contains the source and neighboring entities of the candidate news, and the entities in the historical click news are highlighted in yellow. Fig. 5 shows the attention of a query to neighboring nodes, and in Fig. 4, the five nodes with the highest attention are highlighted in blue. Although this candidate news has a relatively low correlation with the user's historical click news, it has a large number of adjacent nodes with a high correlation with the historical click news. The historical news sequence contains a number of U.S. locations and celebrities, and the candidate news also includes many of these entities in its adjacent nodes. This shows that the candidate news has a deep potential connection with the historical click news, and the matching degree has been greatly improved after considering the neighbor nodes.



Winners and losers from Week 8 in college football led by Wisconsin, Oregon There were highs and lows...

Dr. Maya Rockeymoore Cummings, the widow of the late Rep. Elijah Cummings, appeared to respond to President Trump's previous ...

This waterfront city is the top place in America to retire Plus, 4 affordable alternatives on the beach

Joe Biden reportedly denied Communion at a South Carolina church because of his stance on abortion ...

Did Meghan Markle make a change to her hair? Royal watchers noticed she looked a little different.

[Meghan, Duchess of Sussex]

[Time 100, multiracial people , Canoga Park , Thomas Markle, Doria Ragland, Prince Harry, Duke of Sussex, United States of America, wedding of Prince Harry and Meghan Markle, English, British royal family, Toronto, Los Angeles, Montecito, Frogmore Cottage, Suits, Meghan, Duchess of Sussex]

Figure 4: Historical click sequence and candidate news for the sample user

LLM and KG Augmented News Encoder considers potential connections between candidate news and user's history of clicking on news, which makes user-candidate news matching beyond the understanding of news text.

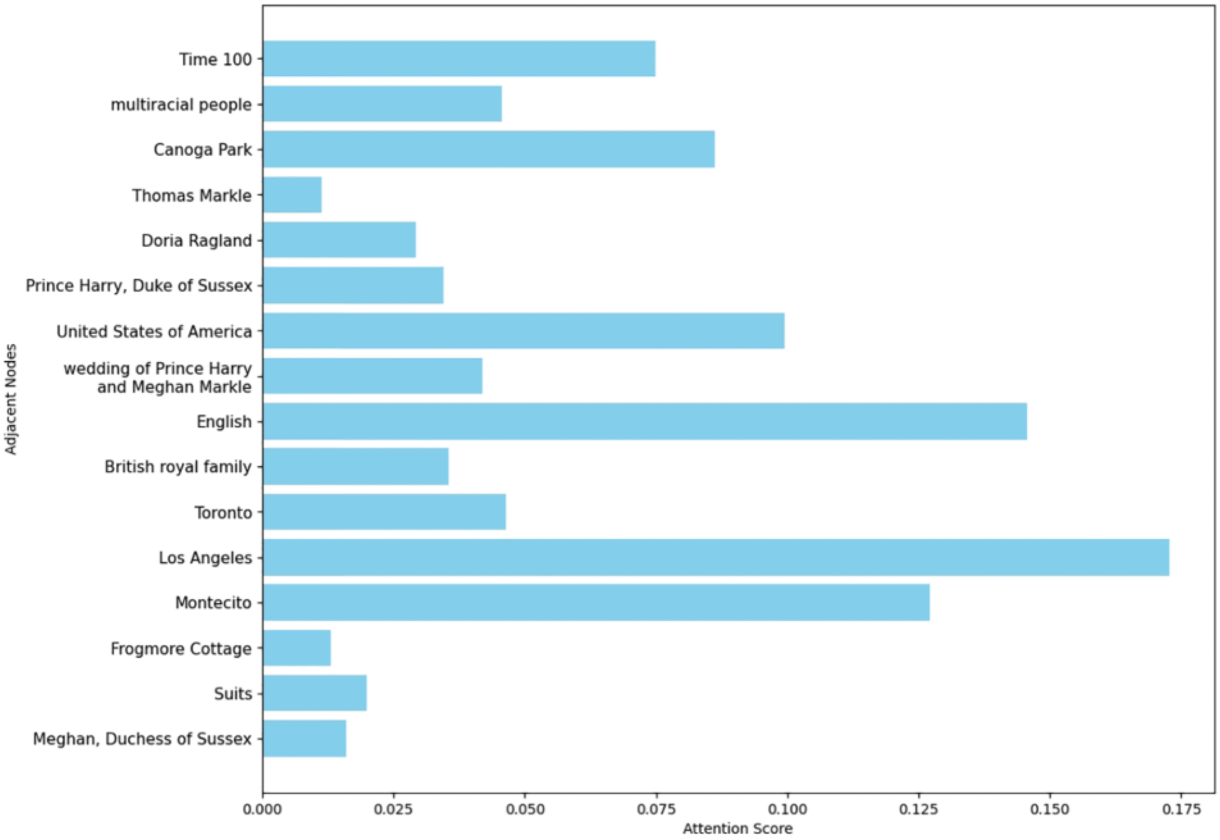


Figure 5: The visualization of the attention weights to adjacent nodes

7 Conclusion

In this work, we propose an innovative personalized news recommendation framework LKPNR, which integrates a Large Language Model (LLM) and a Knowledge Graph (KG). While combining the General News Encoder, the robust contextual comprehension capability of the LLM allows us to derive news representations imbued with semantic information. Simultaneously, we harness the news relationship graph structure inherent in the KG to extract supplementary collaborative news information, enhancing the efficacy of the news recommendation system and alleviating the long tail problem to a certain extent. The experimental results demonstrate the outstanding performance of our proposed framework, leading to significant enhancements over the traditional baseline.

Acknowledgement: The authors would like to thank the editors and reviewers for their valuable work, as well as the colleagues and leaders for their valuable support during the research process.

Funding Statement: The current work was supported by National Key R&D Program of China (2022QY2000-02).

Author Contributions: The authors confirm contribution to the paper as follows: Study conception and design: Zhanwei Xuan, Kai Zhang; data collection: Zhou Yan, Xin Wang; analysis and interpretation of results: Xiangyang Cui, Hao Chen, Runfeng Xie; draft manuscript preparation: Zhanwei Xuan, Hao Chen, Runfeng Xie. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available in https://github.com/Xuan-ZW/LKPNR.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] C. Wu, F. Wu, Y. Huang, and X. Xie, "Personalized news recommendation: Methods and challenges," *ACM Trans. Inf. Syst.*, vol. 41, no. 1, pp. 1–50, 2023. doi: 10.1145/3530257.

[2] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013. doi: 10.1109/TPAMI.2013.50.

[3] F. Wu *et al.*, "MIND: A large-scale dataset for news recommendation," in *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3597–3606.

[4] C. Wu, F. Wu, M. An, J. Huang, Y. Huang and X. Xie, "Neural news recommendation with attentive multi-view learning," in *Proc. Twenty-Eighth Int. Joint Conf. on Artificial Intelligence*, 2019, pp. 3863–3869.

[5] C. Wu, F. Wu, S. Ge, T. Qi, Y. Huang and X. Xie, "Neural news recommendation with multi-head self-attention," in *Proc. 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 6389–6394.

[6] K. Alex, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional networks," in *Proc. 25th Int. Conf. on Neural Information Processing System*, 2012, vol. 1, pp. 1097–1105.

[7] A. Graves, "Long short-term memory," *Supervised Sequence Labelling with Recurrent Neural Networks*, vol. 385, pp. 37–45, 2012.

[8] A. Zeng *et al.*, "GLM-130B: An open bilingual pre-trained model," in *The Eleventh Int. Conf. on Learning Representations*, Kigali, Rwanda, 2022.

[9] H. Touvron *et al.*, "LLaMA: Open and efficient foundation language models," arXiv preprint arXiv:2302.13971, 2023.

[10] B. Peng *et al.*, "RWKV: Reinventing RNNs for the transformer era," arXiv preprint arXiv:2305.13048, 2023.

[11] H. Yin, B. Cui, J. Li, J. Yao, and C. Chen, "Challenging the long tail recommendation," in *Proc. VLDB Endowment*, vol. 5, no. 9, pp. 896–907, 2012. doi: 10.14778/2311906.2311916.

[12] Q. Guo *et al.*, "A survey on knowledge graph-based recommender systems," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 8, pp. 3549–3568, 2020. doi: 10.1109/TKDE.2020.3028705.

[13] H. Wang, M. Zhao, X. Xie, W. Li, and M. Guo, "Knowledge graph convolutional networks for recommender systems," in *The World Wide Web Conf.*, San Francisco, CA, USA, 2019, pp. 3307–3313.

[14] W. Ouyang *et al.*, "Learning graph meta embeddings for cold-start ads in click-through rate prediction," in *Proc. 44th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, Canada, 2021, pp. 1157–1166.

[15] S. Okura, Y. Tagami, S. Ono, and A. Tajima, "Embedding-based news recommendation for millions of users," in *Proc. 23rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Halifax, NS, Canada, 2017, pp. 1933–1942.

[16] J. Lian, F. Zhang, X. Xie, and G. Sun, "Towards better representation learning for personalized news recommendation: A multi-channel deep fusion approach," in *Proc. Twenty-Seventh Int. Joint Conf. Artif. Intell.*, Stockholm, Sweden, 2018, pp. 3805–3811.

[17] M. An, F. Wu, C. Wu, K. Zhang, Z. Liu and X. Xie, "Neural news recommendation with long-and short-term user representations," in *Proc. 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019, pp. 336–345.

[18] A. Vaswani *et al.*, “Attention is all you need,” *Adv. Neural. Inf. Process. Syst.*, vol. 30, pp. 5998–6008, 2017.

[19] C. Wu, F. Wu, M. An, J. Huang, Y. Huang and X. Xie, “NPA: Neural news recommendation with personalized attention,” in *Proc. 25th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*, Anchorage, AK, USA, 2019, pp. 2576–2584.

[20] H. Wang, F. Wu, Z. Liu, and X. Xie, “Fine-grained interest matching for neural news recommendation,” in *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 836–845.

[21] W. C. Kang *et al.*, “Do LLMs understand user preferences? Evaluating LLMs on user rating prediction,” arXiv preprint arXiv:2305.06474, 2023.

[22] Y. Hou *et al.*, “Large language models are zero-shot rankers for recommender systems,” arXiv preprint arXiv:2305.08845, 2023.

[23] Y. Gao, T. Sheng, Y. Xiang, Y. Xiong, H. Wang and J. Zhang, “Chat-REC: Towards interactive and explainable LLMs-augmented recommender system,” arXiv preprint arXiv:2303.14524, 2023.

[24] Q. Liu, N. Chen, T. Sakai, and X. M. Wu, “A first look at LLM-powered generative news recommendation,” arXiv preprint arXiv:2305.06566, 2023.

[25] R. Li, W. Deng, Y. Cheng, Z. Yuan, J. Zhang and F. Yuan, “Exploring the upper limits of text-based collaborative filtering using large language models: Discoveries and insights,” arXiv preprint arXiv:2305.11700, 2023.

[26] H. Wang, F. Zhang, X. Xie, and M. Guo, “DKN: Deep knowledge-aware network for news recommendation,” in *Proc. 2018 World Wide Web Conf.*, Lyon, France, 2018, pp. 1835–1844.

[27] Z. Mao, X. Zeng, and K. F. Wong, “Neural news recommendation with collaborative news encoding and structural user encoding,” in *Findings of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic, 2021, pp. 46–55.

[28] B. Yang, D. Liu, T. Suzumura, R. Dong, and I. Li, “Going beyond local: Global graph-enhanced personalized news recommendations,” in *Proc. 17th ACM Conf. on Recommender Systems*, Singapore, 2023, pp. 24–34.

[29] Y. Sun, Q. Shi, L. Qi, and Y. Zhang, “JointLK: Joint reasoning with language models and knowledge graphs for commonsense question answering,” in *Proc. 2022 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, USA, 2022, pp. 5049–5060.

[30] Y. Ding, B. Wang, X. Cui, and M. Xu, “Popularity prediction with semantic retrieval for news recommendation,” *Expert Syst. Appl.*, vol. 247, pp. 123308, 2024. doi: 10.1016/j.eswa.2024.123308.