HDNR: A Hyperbolic-Based Debiased Approach for Personalized News Recommendation

Shicheng Wang
Institute of Information Engineering, Chinese Academy of Sciences
School of Cyber Security, University of Chinese Academy of Sciences
wangshicheng@iie.ac.cn

Shu Guo*
National Computer Network Emergency Response Technical Team/Coordination Center
guoshu@cert.org.cn

Lihong Wang
National Computer Network Emergency Response Technical Team/Coordination Center
wlh@isc.org.cn

Tingwen Liu
Institute of Information Engineering, Chinese Academy of Sciences
School of Cyber Security, University of Chinese Academy of Sciences
liutingwen@iie.ac.cn

Hongbo Xu
Institute of Information Engineering, Chinese Academy of Sciences
School of Cyber Security, University of Chinese Academy of Sciences
hbxu@iie.ac.cn

ABSTRACT

Personalized news recommendation aims to recommend candidate news to the target user, according to the clicked news history. The user-news interaction data exhibits power-law distribution, however, existing works usually learn representations in Euclidean space which makes inconsistent capacities between data space and embedding space, leading to severe representation distortion problem. Besides, the existence of conformity bias, a potential cause of power-law distribution, may introduce biased guidance to learn user representations. In this paper, we propose a novel debiased method based on hyperbolic space, named HDNR, to tackle the above problems. Specifically, first, we employ hyperboloid model with exponential growth capacity to conduct user and news modeling, in order to solve inconsistent space capacities problem and obtain low distortion representations. Second, we design a re-weighting aggregation module to further mitigate conformity bias in data distribution, through considering local importance of the clicked news among contextual history and its global popularity degree simultaneously. Finally, we calculate the relevance score between target user and candidate news representations. We conduct experiments on two real-world news recommendation datasets MIND-Large, MIND-Small and empirical results demonstrate the effectiveness of our approach from multiple perspectives.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; **Personalization**.

*Corresponding author.

ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3591693

KEYWORDS

news recommendation, hyperbolic space, conformity bias

ACM Reference Format:
Shicheng Wang, Shu Guo, Lihong Wang, Tingwen Liu, and Hongbo Xu. 2023. HDNR: A Hyperbolic-Based Debiased Approach for Personalized News Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3539618.3591693

1 INTRODUCTION

Recently, with the rapid development of Internet and online news services [1, 2], massive news are released on numerous news platforms all the time which makes users overwhelmed. Hence, personalized news recommendation is necessary for news platforms to help users alleviate information overload as well as improve their reading news experience. Briefly, news recommendation techniques aim to recommend candidate news that target user may be interested in.

Current news recommendation researches can be divided into ID-based methods, content-based methods and hybrid methods. Traditional ID-based collaborative filtering methods [1, 3] reconstruct interactive matrix to learn user and news representations. However, they suffer from severe cold-start problem due to the short life cycles characteristic of news articles, leading to explore content-based methods. The content-based methods are designed to learn news semantic information and explore users' reading interests accurately. Generally speaking, they share a universal framework, including news encoder, user encoder and click predictor. In the light of this framework, existing methods usually adopt Transformer [4] or BERT [5] to learn news representations based on text content such as news titles [6, 7]. User interest modeling is another significant procedure in this framework. With the extensive application of deep learning techniques, Recurrent Neural Network and Transformer, regarded as effective methods for dealing with sequence modeling, are adopted to identify users' reading interests [8–10]. Due to the huge influence of graph neural networks techniques, some works introduce Graph Attention Networks [11]

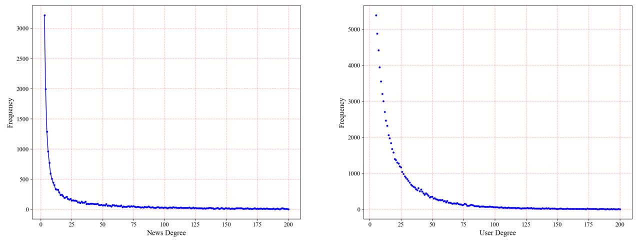

Figure 1: Degree distributions of user and news in MIND-Small Dataset.

into user modeling process to enhance users representations by leveraging information propagation and aggregation [12, 13].

We conduct statistic analysis on real-world dataset MIND-Small and observe that the user-news bipartite network exhibits power-law distribution, as shown in Figure 1, which means a minority of popular news account for the majority of interaction behaviors. In particular, it has been proved theoretically [14] that such scale-free graph can be traced back to hierarchical structure and the graph volume grows exponentially with its radius. However, existing news recommendation methods usually tend to learn user and news semantic representations in Euclidean space owning polynomial growth volume with respect to radius. As a result, the inconsistent capacities between data space and embedding space leads to severe representation distortion problem. Besides, the widespread existence of conformity bias, which is a potential cause of power-law distribution, aggravates the imbalance of data distribution. Since conformity bias happens as users tend to behave similarly to the others even if the behaviors against their own judgment [15], users' interactive history perhaps cannot reflect their intrinsic interests accurately, especially when clicking on popular news. Nevertheless, recent works usually ignore the conformity bias, resulting in introducing biased guidance to the recommendation model and producing negative effect.

To tackle the above-mentioned problems, in this work we propose a novel debiased method based on hyperbolic space for news recommendation, termed as HDNR. Specifically, first, we employ hyperboloid model with exponential growth volume to conduct user and news modeling, in order to obtain low distortion representations when embedding power-law distribution data with hierarchical structures. Second, on the basis of traditional user modeling, we design a re-weighting aggregation module considering local relative importance of the clicked news among contextual history and its global popularity degree simultaneously. In this way, we are able to mitigate conformity bias in data distribution as well as emphasize users' intrinsic interests, producing more accurate users representations ultimately. Finally, we calculate the relevance score between target user and candidate news representations and decide whether to recommend or not. We conduct experiments on real-world datasets MIND-Large and MIND-Small to demonstrate the effectiveness of our proposed HDNR.

To summarize, the major contributions of this paper include:

- We utilize the superiority of hyperbolic space to conduct user and news modeling, in order to solve inconsistent space capacities problem and obtain low distortion representations.
- We design a novel re-weighting aggregation module to mitigate conformity bias in data distribution as well as emphasize users' intrinsic interests, through considering local relative importance of the clicked news among history and its global popularity degree simultaneously.
- The empirical results demonstrate the effectiveness of our approach and verify the validity of hyperbolic geometry as well as re-weighting aggregation module.

2 RELATED WORK

2.1 Personalized News Recommendation

With the growth of individual and social needs, news recommendation has attracted more and more attention recently. Therefore, a variety of methods have been proposed, including ID-based methods, content-based methods, and hybrid methods. Most traditional ID-based methods achieved news recommendation based on collaborative filtering framework [1, 3]. They parameterized users and items in a latent space and aimed at reconstructing interactive behaviors. However, due to the short life cycles characteristic of news articles, collaborative filtering methods based on IDs always suffered from severe cold start problem, which required us to understand news contents and user interests.

To this end, content-based methods and hybrid methods have been proposed, which treat the task as a sequence modeling problem in the early stage. For example, Okura [16] utilized an auto-encoder to learn news representations from text content. Then they applied a GRU network to model user interests from clicked history. NAML [9] leveraged a CNN network to learn news semantic representations from news titles and categories. Then they learnt user representations through attentively aggregating clicked history. NRMS [7] utilized homologous multi-head self-attention networks to learn news representations and user representations separately, capturing interactive information among word sequences and news sequences respectively. Then with the development of graph neural network techniques, GNN-based methods are proposed to learn news and user representations based on distinct interaction graphs, due to the advantages of aggregating high-order neighborhood information. For example, KRED [17] first introduced news titles and entities to construct a news recommendation knowledge graph. Then they applied graph attention network to learn news representations. GNewsRec [13] learnt users' short-term interests by applying attentive GRU neural network to clicked history as well as users' long-term interests via graph neural networks.

However, existing work usually tend to learn user and news representations in Euclidean space, ignoring that the flat polynomial growth space is incompatible with the intrinsic hierarchical structure. Different from these methods, in this paper, we resort to hyperbolic space to obtain low distortion representations, resulting in better recommendation performance.

2.2 Hyperbolic Neural Networks

Since the latent hierarchical structure is a generic property of real-world data, massive researches have explored the potential of hyperbolic space in different domain, such as computer vision, natural language processing and graph representation learning, etc. For instance, Nickel et al. [18] proposes to learn representations for symbolic data by embedding them into hyperbolic space, while considering the latent hierarchical structure. Chami et al. [19] leverages both the expressiveness of GCN and hyperbolic geometry to learn node representations for scale-free graphs. Gulcehre et al. [20] extends attention mechanism into hyperbolic space and shows improvements in terms of many NLP tasks including neural machine translation and visual question answering.

Recently, considering the prevalence of the power-law distribution in user-item bipartite networks, hyperbolic geometric has attracted a lot of attention and been applied to recommender systems. For example, HGCF [21] incorporates multiple layers of neighborhood aggregation using a hyperbolic GCN module to obtain higher-order information and the underlying hierarchical structure in user-item interactions. HRCF [22] proposes a geometric-aware hyperbolic regularizer for superior performance regarding the geometric properties of hyperbolic space. HICF [23] aims to make the pull and push components of the hyperbolic margin ranking loss to be geometric aware, instead of simply migrating the loss function of Euclidean space to hyperbolic space. LKGR [24] presents a knowledge-aware attention mechanism on the Lorentzian manifold and stack multiple layers to aggregate high-order information for knowledge-aware recommendation. HSR [25] designs a hyperbolic aggregator on the users' social neighbors to take full advantage of the social information and enhance the Performance in social recommendation. HyperSoRc [26] exploits the hyperbolic user and item presentations with multiple social relations.

However, despite the superiority of hyperbolic space, there is still huge room for improvement since the majority of the current works merely generalize Euclidean models to hyperbolic space.

3 PRELIMINARIES

Hyperbolic geometry is a non-Euclidean geometry with a constant negative curvature, which measures how a geometric object deviates from a flat plane. In this paper we adopt hyperboloid model, one of the typical equivalent hyperbolic models, due to its simplicity and numerical stability.

Definition 1 (Hyperboloid Manifold). An d-dimensional hyperboloid model (alternatively called Lorentz model) with constant negative curvature $-1/K(K > 0)$ is a manifold embedded in the $d+1$ dimensional Minkowski space, denoted as $\mathbb{H}^{d,K}$, where $\mathbb{H}^{d,K} = \{\mathbf{x} \in \mathbb{R}^{d+1} :< \mathbf{x}, \mathbf{x} >_{\mathbb{H}}= -K, x_0 > 0\}$ and $< .,. >_{\mathbb{H}}$ represents the Lorentz inner product operation in the hyperbolic space. Given $\mathbf{x}, \mathbf{y} \in \mathbb{H}^{d,K}$, the Lorentz inner product is defined as follows:

$$< \mathbf{x}, \mathbf{y} >_{\mathbb{H}}= -x_0 y_0 + x_1 y_1 + \cdots + x_d y_d. \tag{1}$$

In addtion, the d-dimensional Euclidean tangent space centered at vector $\mathbf{x} \in \mathbb{H}^{d,K}$ is denoted as $\mathbb{T}^{d,K}_{\mathbf{x}}$, where $\mathbb{T}^{d,K}_{\mathbf{x}} = \{\mathbf{v} \in \mathbb{R}^{d+1} :< \mathbf{v}, \mathbf{x} >_{\mathbb{H}}= 0\}$. In other words, $\mathbb{T}^{d,K}_{\mathbf{x}}$ is defined as the first-order approximation of the hyperboloid manifold $\mathbb{H}^{d,K}$ around x.

Definition 2 (Exponential and Logarithmic Mappings). Exponential and Logarithmic mappings are proposed to bridge the hyperbolic space and tangent space. Given $\mathbf{x}, \mathbf{y} \in \mathbb{H}^{d,K}$ and $\mathbf{v} \in \mathbb{T}^{d,K}_{\mathbf{x}}$, the exponential mapping maps $\mathbf{v}$ to the hyperbolic space, mathematically defined as:

$$\exp^K_{\mathbf{x}}(\mathbf{v}) = \cosh(\frac{||\mathbf{v}||_{\mathbb{H}}}{\sqrt{K}})\mathbf{x} + \sqrt{K}\sinh(\frac{||\mathbf{v}||_{\mathbb{H}}}{\sqrt{K}})\frac{\mathbf{v}}{||\mathbf{v}||_{\mathbb{H}}}. \tag{2}$$

The logarithmic mapping, the inverse operation of $\exp^K_{\mathbf{x}}$, is defined as:

$$\log^K_{\mathbf{x}}(\mathbf{y}) = d^K_{\mathbb{H}}(\mathbf{x}, \mathbf{y})\frac{\mathbf{y} + \frac{1}{K} < \mathbf{x}, \mathbf{y} >_{\mathbb{H}} \mathbf{x}}{||\mathbf{y} + \frac{1}{K} < \mathbf{x}, \mathbf{y} >_{\mathbb{H}} \mathbf{x}||_{\mathbb{H}}}, \tag{3}$$

where $d^K_{\mathbb{H}}(\cdot, \cdot)$ is the distance between two points $\mathbf{x}, \mathbf{y} \in \mathbb{H}^{d,K}$ on hyperboloid manifold, which is formulated as:

$$d^K_{\mathbb{H}}(\mathbf{x}, \mathbf{y}) = \sqrt{K}\text{arcosh}(-\frac{< \mathbf{x}, \mathbf{y} >_{\mathbb{H}}}{K}). \tag{4}$$

For simplicity, we fix parameter K and set it to 1, implying that the curvature is -1.

4 METHODOLOGY

In this section, we first present the problem formulation of personalized news recommendation.

Task Definition. Given a candidate news n_c and target user u with his clicked history $[n_1, \cdots, n_N]$, we aim to learn candidate news representation as well as user representation respectively. Afterwards we calculate the relevance score between their representations and decide whether to recommend n_c or not according to the score.

Then we propose a hyperbolic-based debiased method, named HDNR, in order to learn low distortion representations as well as eliminate conformity bias in data distribution. Figure 2 illustrates the overall framework of our HDNR model. Generally, it consists of three components: News Encoder, Hyperbolic User Encoder, and Click Predictor. Within hyperbolic user encoder, a novel re-weighting aggregation module is designed to modify users representations obtained from fundamental hyperbolic attention module. We will elaborate our method in the subsequent sections in detail.

4.1 News Encoder

In this section, we introduce how to generate initial news representations. We first learn semantic representations from news titles on Euclidean space. Given the title word sequence $[w_1, \cdots, w_M]$, where w_i is denoted as the i-th word in title, we utilize the traditional Transformer framework to encode titles.

Specifically, at the bottom of news encoder module, it applies word embedding layer to embed each word w_i into corresponding vector $e_i \in \mathbb{R}^d$. Then, it adopts a multi-head self-attention network to capture word-level semantic relatedness among entire words $[e_1, \cdots, e_M]$.

$$\alpha^s_{i,j} = \frac{\exp(e_i^T Q^w_s e_j)}{\sum_{t=1}^M \exp(e_i^T Q^w_s e_t)}. \tag{5}$$

Then the representation of the word w_i learned by the s-th attention head h^s_i is calculated as:

$$h^s_i = V^w_s \sum_{j=1}^M \alpha^s_{i,j} \cdot e_j, \tag{6}$$

where $Q^w_s \in \mathbb{R}^{d \times d}$ and $V^w_s \in \mathbb{R}^{d/S \times d}$ are the projection parameters in the s-th self-attention head, and $\alpha_{i,j}$ indicates the interaction score between the word w_i and w_j. Then the multi-head representation of word w_i is concatenated as $h_i \in \mathbb{R}^d$, i.e., $h_i = [h^1_i; h^2_i; \cdots; h^S_i]$, where S denotes the number of separate self-attention heads. Next it applies an additive attention network to aggregate

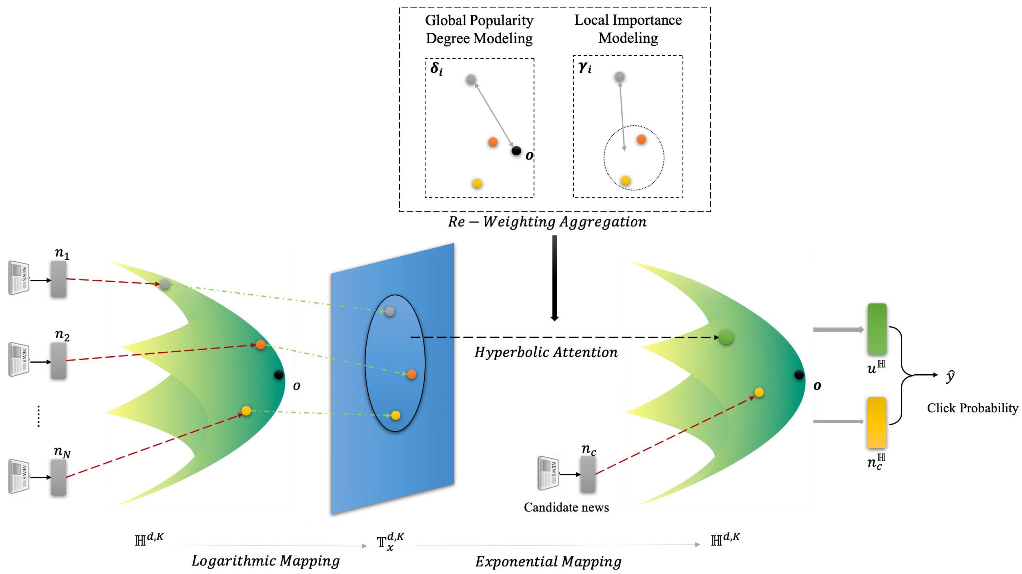

Figure 2: Overall framework of our HDNR method.

contextual word representations into an Euclidean news representation n, formulated as:

$$\alpha_i = \frac{\exp(q_\alpha^T \cdot h_i)}{\sum_{j=1}^{M} \exp(q_\alpha^T \cdot h_j)}, \tag{7}$$

$$n^{\mathbb{E}} = \sum_{i=1}^{M} \alpha_i \cdot h_i, \tag{8}$$

where $q_\alpha \in \mathbb{R}^d$ is a projection parameter vector, and α_i denotes the contribution weight of word w_i to news representation.

Since the user-news bipartite network exhibits power-law distribution, we need to explicitly encode Euclidean news representations onto the hyperboloid manifold via exponential mapping. Let $\mathbf{o} := \{\sqrt{K}, 0, \cdots, 0\} \in \mathbb{H}^{d,K}$ denote the north pole (origin) in $\mathbb{H}^{d,K}$, which we use as a reference point to perform mapping transformations. Since it is obvious that $<(0, \mathbf{x}^{\mathbb{E}}), \mathbf{o}>= 0$, we interpret $(0, \mathbf{x}^{\mathbb{E}})$ as a point in tangent space $\mathbb{T}^{d,K}$. Therefore, the low distortion hyperbolic news representations can be encoderd as follows:

$$n^{\mathbb{H}} = \exp_{\mathbf{o}}^K((0, n^{\mathbb{E}})) = (\sqrt{K}\cosh(\frac{||n^{\mathbb{E}}||_2}{\sqrt{K}}), \sqrt{K}\sinh(\frac{||n^{\mathbb{E}}||_2}{\sqrt{K}})\frac{n^{\mathbb{E}}}{||n^{\mathbb{E}}||_2}). \tag{9}$$

4.2 Hyperbolic User Encoder

4.2.1 Hyperbolic Attention Module.
In this section, we design hyperbolic attention mechanism to aggregate hyperbolic initial news representations on hyperboloid manifold, so as to obtain users' interests representations. Since operations on hyperboloid manifold tend to be complicated, we resort to exponential and logarithmic mappings to achieve information aggregation.

At the beginning, since there is semantic relatedness between news articles clicked by the same user, we project news representations onto tangent space and then apply a multi-head self-attention to enhance the news representations by capturing news-level relatedness among clicked news history $[n_1^{\mathbb{H}}, \cdots, n_N^{\mathbb{H}}]$,

$$\beta_{i,j}^s = \frac{\exp((\log_{\mathbf{o}}^K(n_i^{\mathbb{H}}))^T Q_s^n \log_{\mathbf{o}}^K(n_j^{\mathbb{H}}))}{\sum_{t=1}^{N} \exp((\log_{\mathbf{o}}^K(n_i^{\mathbb{H}}))^T Q_s^n \log_{\mathbf{o}}^K(n_t^{\mathbb{H}}))}. \tag{10}$$

Thus the news representation learned by the s-th attention head n_i^s is formulated as:

$$n_i^s = V_s^n \sum_{j=1}^{N} \beta_{i,j}^s \cdot \log_{\mathbf{o}}^K(n_j^{\mathbb{H}}), \tag{11}$$

where $Q_s^n \in \mathbb{R}^{d\times d}$ and $V_s^n \in \mathbb{R}^{d/S\times d}$ are the news-level projection parameters in the s-th self-attention head. Then the multi-head representation of news n_i is concatenated, followed by exponential mapping operation, i.e., $n_i^{\mathbb{H}} = \exp_{\mathbf{o}}^K([n_i^1; n_i^2; \cdots; n_i^S])$.

Similarly, we apply another attention network followed by the exponential mapping to aggregate historical news information into a hyperbolic user representation, formulated as:

$$\beta_i = \frac{\exp(q_\beta^T \cdot \log_{\mathbf{o}}^K(n_i^{\mathbb{H}}))}{\sum_{j=1}^{N} \exp(q_\beta^T \cdot \log_{\mathbf{o}}^K(n_j^{\mathbb{H}}))}, \tag{12}$$

$$u^{\mathbb{H}} = \exp_{\mathbf{o}}^K(\sum_{i=1}^{N} \beta_i \cdot \log_{\mathbf{o}}^K(n_i^{\mathbb{H}})), \tag{13}$$

where $q_\beta \in \mathbb{R}^d$ is a projection parameter vector. In this way, we are able to obtain low distortion user representations through utilizing hyperbolic attention mechanism. However, due to the widespread existence of conformity bias in data distribution, users' historical behaviors cannot reflect their intrinsic interests accurately. Therefore, we expect to learn personalized user representations to emphasis their intrinsic reading interests and eliminate conformity bias as much as possible.

4.2.2 Re-Weighting Aggregation Module.

Considering the problems identified in the previous section, we investigate the information aggregation procedure and argue that the attention weight β_i ought to be more personalized. Intuitively, on the one hand, the weights for clicked news that users exactly interested in are supposed to increase; on the other hand, the weights for clicked news caused by conformity psychology should be reduced. To this end, we propose a re-weighting aggregation module, including local importance modeling and global popularity degree modeling, to modify the weight β_i, enabling us to obtain more accurate user representations. Here we provide a detailed introduction on two submodules.

Local Importance Modeling aims to describe the relevance between each clicked news and the target user. Since we are unaware of the users' intrinsic interests, the objective is approximately calculated as the relevance score between each clicked news and the entire contextual history, formulated as follows:

$$\gamma_i = \sum_{j=1}^{N} (\log_{\mathbf{o}}^{K}(n_i^{\mathbb{H}}))^T \cdot W_\gamma \cdot \log_{\mathbf{o}}^{K}(n_j^{\mathbb{H}}), \tag{14}$$

where $W_\gamma \in \mathbb{R}^{d\times d}$ is a projection matrix. That is, the larger γ_i is, the more significant n_i is among the target user's historical records. In other words, γ_i tends to endow higher weights to the news that reflect users' real reading interests.

Global Popularity Degree Modeling is expected to reduce the weights of clicked news caused by conformity psychology. Although it is difficult to distinguish the genuine reasons for clicking, the conformity phenomenon is more likely to happen when clicking on popular news rather than non-popular ones. Thus we put forward an approximate solution utilizing global structural properties, whose purpose is to decrease the weights of popular news and enlarge the weights of the others, calculated as follows:

$$\delta_i = d_{\mathbb{H}}^{K}(\exp_{\mathbf{o}}^{K}(W_\delta \cdot \log_{\mathbf{o}}^{K}(n_i^{\mathbb{H}})), \mathbf{o}), \tag{15}$$

where $W_\delta \in \mathbb{R}^{d\times d}$ is a projection matrix. Due to the properties of power-law distribution itself, popular news are arranged near the north pole (origin) when embedded in hyperbolic space while non-popular news are placed around the boundary. In this way, as the distance on hyperboloid manifold decreases, the weight δ_i decreases too, which is able to mitigate conformity bias in data distribution to a certain extent.

Afterwards, on the basis of original attention weight β_i, we incorporate the local importance weight γ_i and global popularity degree weight δ_i together, achieving the two-fold purposes through the comprehensive effect of the proposed submodules. he modified attention weight $\beta_i^{'}$ is calculated as:

$$\beta_i^{'} = SoftMax(\beta_i \cdot \gamma_i \cdot \delta_i). \tag{16}$$

Finally, the modified news representations are aggregated to generate more accurate user representations, which are more likely to reflect users' intrinsic interests:

$$u^{\mathbb{H}} = \exp_{\mathbf{o}}^{K}(\sum_{i=1}^{M} \beta_i^{'} \cdot \log_{\mathbf{o}}^{K}(n_i^{\mathbb{H}})). \tag{17}$$

Notably, multiple layers can be stacked to aggregate high-order relatedness information to enhance user and news representations, which we do not implement this time due to resource limitation. We will further explore the implementation in the future.

Table 1: Statistic information of MIND-Large dataset.

# News	161,013	# Users	1,000,000
# News category	20	# Impression	15,777,377
# Entity	3,299,687	# Click behavior	24,155,470
Avg. title len.	11.52	Avg. abstract len.	43.00
Avg. body len.	585.05		

4.3 Click Predictor

Simple but effective, the click probability score y is computed by the inner product between the target user representation $u^{\mathbb{H}}$ and the candidate news representation $n_c^{\mathbb{H}}$, i.e.,

$$\hat{y} = \log_{\mathbf{o}}^{K}(u^{\mathbb{H}}) \cdot \log_{\mathbf{o}}^{K}(n_c^{\mathbb{H}}). \tag{18}$$

During the evaluation stage, candidate news with top scores y are selected to be recommended to the target user.

4.4 Model Training

Following [7], we use negative sampling techniques for model training. Denoting clicked candidate news in the training set as positive sample n_i, i.e., $y_{u,i} = 1$, then we randomly choose P non-clicked candidate news as negative samples $[n_{i,1}, \cdots, n_{i,P}]$ from the same impression displayed to the target user. The corresponding click probability scores of the positive and P negative samples are calculated as $\hat{y}_i^+$ and $[\hat{y}_{i,1}^-, \hat{y}_{i,2}^-, \cdots, \hat{y}_{i,P}^-]$ respectively. Finally, the supervised classification loss function is formulated as follows:

$$\mathcal{L}_{ce} = -\sum_{i\in TS} \log \frac{\exp(\hat{y}_i^+)}{\exp(\hat{y}_i^+) + \sum_{j=1}^{P} \exp(\hat{y}_{i,j}^-)}, \tag{19}$$

where TS denotes all positive samples.

5 EXPERIMENT

This section conducts experiments to evaluate the performance of our model, HDNR.

5.1 Dataset and Evaluation Metrics

We conduct extensive experiments on two large-scale real-world datasets, MIND-Large [1] and MIND-Small [2], to evaluate the effectiveness of our method. MIND-Large dataset collected from Microsoft News platform contains two record documents. One document describes text content of news, including titles and abstracts. The other document describes interaction behaviors between users and news. These total click behaviors are gathered from October 12 to

[1]https://msnews.github.io/

[2]A small version of the MIND-Large dataset by randomly sampling 50,000 users and their behavior logs.

November 22, 2019 (six weeks). The click behaviors in the first four weeks are regarded as user reading history, the behaviors in the penultimate week is applied for training, and the data in last week is used for performance evaluation. Detailed statistic information about MIND-Large dataset is summarized in Table 1. Following [7], we use four classical metrics, i.e., AUC, MRR [27], nDCG@5 [28], and nDCG@10, for performance evaluation. Notably, AUC is the most important one among them.

5.2 Baselines

To demonstrate the effectiveness of the proposed news recommendation method, we compare our HDNR with two types of the prominent content-based baselines: Sequence-based methods (EBNR, DKN, NPA, NAML, LSTUR, NRMS), and GNN-based methods (GNewsRec, GERL, User-as-Graph) , as follows [3].

(1) **EBMR** [16] employs a GRU network to learn user representations from clicked news history.

(2) **DKN** [29] utilizes an adaptive attention network to learn user representations considering relatedness between candidate news and historical news.

(3) **NPA** [10] employs personalized attention networks to learn individual representation for each user.

(4) **NAML** [9] leverages CNN networks to model news semantic representations and learns user representations through attentively aggregating clicked news.

(5) **LSTUR** [6] models short-term user interests via a GRU network and long-term user interests via user ID embeddings respectively from two perspectives.

(6) **NRMS** [7] learns news representations and user representations through utilizing multi-head self-attention networks respectively.

(7) **GNewsRec** [13] models user short-term interests by applying attentive GRU neural network and user long-term interests via graph neural networks based on user-news-topic heterogeneous graph.

(8) **GERL** [12] utilizes the neighborhood of news and users on the user-news bipartite network to enhance their representations.

(9) **User-as-Graph** [30] proposes a heterogeneous graph pooling method to learn user interest representations from the personalized heterogeneous graph.

5.3 Implementation Details

Next, we introduce experimental and hyper-parameters settings of our method. For news content modeling, we utilize the first 30 words of news titles to learn corresponding news representations. In addition, a special character [PAD] is used for filling when the length of word sequence does not meet the condition. Besides, we adopt pre-trained Glove embeddings [31] for word initialization. For user interest modeling, we treat the recent 50 clicked news as users' reading history. Moreover, news representations and user representations, as well as latent embeddings are both 400-dimensional vectors, i.e., $d = 400$. For hyper-parameters, the negative sampling ratio is 4. In addition, we utilize dropout technique and Adam optimizer [32] for training. The dropout rate and learning rate are 0.1 and 0.001 respectively.

[3]Due to the limitation of computation resources, we did not employ the pretrained language models to encode news content and report results based on pretrained models.

Table 2: Performance of different methods on MIND-Large Dataset (%). *The improvement is significant at the level p < 0.001.

Method	AUC	MRR	nDCG@5	nDCG@10
EBNR	65.42	31.24	33.76	39.47
DKN	64.60	31.32	33.84	39.48
NPA	66.69	32.24	34.98	40.68
NAML	66.86	32.49	35.24	40.91
LSTUR	67.73	32.77	35.59	41.34
NRMS	67.76	33.05	35.94	41.63
GNewsRec	67.53	32.68	35.46	41.17
GERL	68.24	33.46	36.38	42.11
User-as-Graph	<u>69.23</u>	**34.14**	<u>37.21</u>	<u>43.04</u>
HDNR*	**69.98**	<u>34.10</u>	**37.97**	**44.20**

Table 3: Performance of different methods on MIND-Small Dataset (%). *The improvement is significant at the level p < 0.001.

Method	AUC	MRR	nDCG@5	nDCG@10
EBNR	61.62	28.07	30.55	37.07
DKN	63.99	28.95	31.73	37.07
NPA	64.28	29.64	32.28	38.93
NAML	64.30	29.81	32.64	39.11
LSTUR	65.68	30.44	33.49	39.95
NRMS	65.43	30.74	33.13	39.66
GNewsRec	65.91	30.50	33.56	40.13
GERL	66.22	30.89	34.28	40.50
User-as-Graph	<u>66.71</u>	<u>31.13</u>	<u>34.51</u>	<u>40.95</u>
HDNR*	**68.23**	**32.61**	**36.10**	**42.29**

5.4 Overall Performance

The main purpose of this section is to verify the effectiveness of our method. We conduct experiments to compare our proposed HDNR with several baselines on MIND-Large dataset and then apply them to MIND-Small dataset for supplement. The overall performance results are displayed in Table 2 and Table 3 respectively, where the best results are in bold and the second best are underlined.

Obviously, we have several observations according to the tables: First, our proposed method HDNR outperforms all baselines in terms of AUC on both two real-world datasets, achieving 1.08% and 2.28% improvement comparing to state-of-the-art baselines respectively. Besides, in terms of other evaluation metrics, our method performs excellent as well, which demonstrates the effectiveness and adaptability to data scale of our method. Second, since our method is classified as a sequence-based method, we compare HDNR with the prominent sequence-based baselines. In terms of AUC, our HDNR

Table 4: Performance results on Popular News Oriented Scenario (%).

Method	AUC	MRR	nDCG@5	nDCG@10
NRMS	47.30	25.68	26.84	32.34
User-as-Graph	50.27	24.93	25.77	31.54
HDNR(Ours)	**51.34**	**28.27**	**29.53**	**34.58**
$\Delta_{\mathcal{H}}$	+2.13	+10.09	+10.02	+6.93

Table 5: Performance results on Non-Popular News Oriented Scenario (%).

Method	AUC	MRR	nDCG@5	nDCG@10
NRMS	66.25	31.04	34.22	40.85
User-as-Graph	67.33	32.05	35.23	41.80
HDNR(Ours)	**68.68**	**33.15**	**36.75**	**43.11**
$\Delta_{\mathcal{H}}$	+2.01	+3.43	+4.31	+3.13

obtains remarkable performance improvements over 3% on both datasets. This promising result validates the necessity of our motivations and validity of the proposed methodology. Finally, as shown in tables, GNN-based baselines tend to achieve better performance than sequence-based methods consistently since they aggregate high-order relatedness information from neighborhood additionally to enrich semantic representations. Surprisingly, although we do not adopt GNN components, our HDNR still outperforms all GNN-based baselines, which further demonstrates the superiority of our proposed method.

5.5 Performance on Independent Partitions

As claimed in Introduction section, we recognize that news degree exhibits power-law distribution and popular news tend to attract much more attention due to the widespread existence of conformity bias. To further illustrate the validity of our method, we conduct an in-depth performance comparison on three partitions of the original test dataset. To be specific, we divide the original test data into three independent groups: the popular news oriented scenario (the positive samples in the session are popular news exclusively), the non-popular news oriented scenario (the positive samples in the session are non-popular news exclusively), and hybrid scenario (the positive samples mixed by popular and non-popular news). Experimentally, popular and non-popular news are chosen by the 20/80 rule, also known as Pareto Principle, which can be expressed as a power-law distribution mathematically. In other words, popular news denotes the top 20% of all news sorted by degree, while non-popular news represents the remaining 80%. For simplicity, we compare with two prominent baselines, NRMS and User-as-Graph, representing sequence-based methods as well as GNN-based methods respectively. The Performance results are evaluated on MIND-Small dataset and recorded in Table 4, Table 5 and Table 6 severally, corresponding to different scenarios.

As shown in the tables, we have several following observations: In all partial circumstances, our proposed HDNR outperforms the typical baselines in terms of all evaluation metrics consistently and

Table 6: Performance results on Hybrid Scenario (%).

Method	AUC	MRR	nDCG@5	nDCG@10
NRMS	64.06	14.30	18.20	23.62
User-as-Graph	64.17	13.55	16.93	22.81
HDNR(Ours)	**65.75**	**14.58**	**18.65**	**24.61**
$\Delta_{\mathcal{H}}$	+2.46	+1.96	+2.47	+4.19

obtains at least 2% improvement in terms of AUC regardless of the scenario. The inspiring results suggest that our HDNR model is able to achieve favourable recommending effect whether the candidate news is popular or not, which confirms the universal superiority of our model.

To be specific, the situation that our method provides superior results in the popular news oriented scenario can be mainly attributed to the local importance modeling submodule. As the submodule describes the relatedness between each clicked news and target user in a personalized manner, we can identify the news that target user actually interested in, promoting users' intrinsic interests exploring. Afterwards, when analyzing the non-popular news oriented scenario, in addtion to the local importance modeling submodule, the designed global popularity degree modeling submodule is capable of boosting the weights of non-popular news, preventing them from being suppressed by popular ones. In this way, through utilizing the comprehensive effect of the two submodules, we are able to explore users' real interests as well as mitigate the conformity bias, which makes it possible to recommend non-popular news precisely. The advantages and the consistent improvements on independent scenarios lead to better results on the overall performance.

Although our HDNR obtains better effect than the selected baselines, we should be aware of the huge gap in performance between the popular news oriented scenario and the others. Apparently, the reasons for clicking on popular news are much more complicated than non-popular ones, since predicting the former is perhaps beyond users' intrinsic interests. The relative lower AUC implies that the problems existing in this situation deserve further exploration.

5.6 Ablation Study

In this section, we conduct elaborate ablation experiments on MIND-Small dataset, in order to further evaluate the effectiveness of all innovative components in our method. The results of comparative experiments are recorded in Table 7 in terms of all evaluation metrics. Then we discuss how each component affects the recommendation performance, according to corresponding variant.

Impact of Hyperbolic Geometry. We conduct two groups of contrastive experiments, HDNR versus EDNR and HNNR versus ENNR, to investigate the impact of hyperbolic geometry. Briefly, on the basis of complete model HDNR, EDNR is the variant that Euclidean space substitutes for hyperbolic space. Similarly, the relationship between ENNR and HNNR is identical. Besides, both in the latter group do not adopt re-weighting aggregation module. As shown in Table 7, compared with the results of the models based on hyperbolic space, the corresponding Euclidean space variants droped by 0.83% (68.23 $\rightarrow$ 67.67) and 1.42% (67.76 $\rightarrow$ 66.78) in terms of AUC respectively. In this way, we verify the superiority of

Table 7: Effect of each module perform in our model.

Method	AUC	MRR	nDCG@5	nDCG@10
HDNR*	**68.23**	**32.61**	**36.10**	**42.29**
w/o Hyperbolic Space (abbr. EDNR)	67.67	32.26	35.62	41.95
w/o Re-Weighting (abbr. HNNR)	67.73	32.43	35.94	42.08
w/o Local Importance (abbr. HGNR)	67.48	31.84	35.25	41.64
w/o Global Popularity Degree (abbr. HLNR)	67.42	31.77	35.00	41.55
w/o Hyperbolic Space & Re-Weighting (abbr. ENNR)	66.78	31.17	34.56	40.89

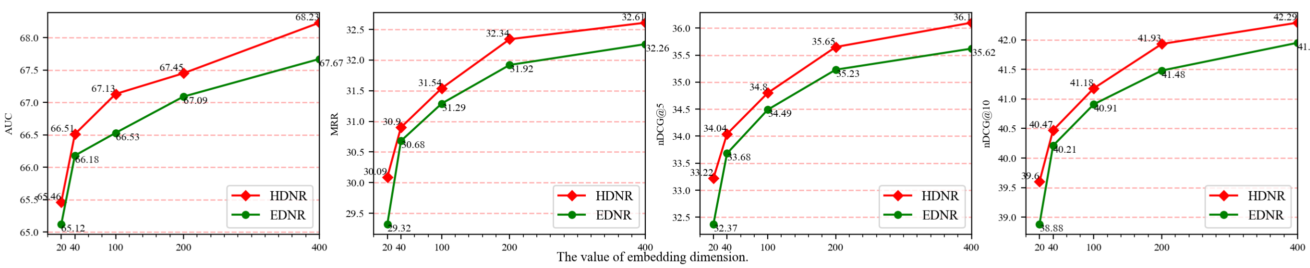

Figure 3: Performance with different embedding dimensions between HDNR and EDNR.

hyperbolic space compared to Euclidean space, enabling us to obtain low distortion representations. In addition, since the decline scope in the former group is not as significant as that in the latter group, we speculate that the effect of hyperbolic geometry is negatively correlated with the complexity of the overall model.

Impact of Re-Weighting Aggregation Module. In order to validate the effectiveness of our re-weighting aggregation module, we construct variants for comparison by removing the proposed module from HDNR and EDNR respectively, namely HNNR and ENNR. As a result of disabling the module, the performance of the latter two shows significant decline of 1.32% (68.23 $\rightarrow$ 67.73) and 1.33% (67.67 $\rightarrow$ 66.78) in terms of AUC severally, which demonstrates the effectiveness of the proposed module. It is noticed that the re-weighting aggregation module consists of two parts. We also examine the impact of each fine-grained submodule. The variant HGNR merely resorts to the global structural properties to reduce the influence of conformity bias in data distribution, while ignoring whether the user is interested in the popular news. Similarly, HLNR only utilizes local relatedness to assist personalized user modeling but neglects the existence of conformity psychology. Obviously, the performance of both incomplete variants decreases dramatically. We suggest that merely introducing either submodule brings negative impact on original hyperbolic aggregation mechanism, i.e., HNNR.

To conclude, we verify the effectiveness of each complete module from multiple perspectives first. Then we observe that both submodules in re-weighting aggregator are indispensable, otherwise leading to unexpected performance degradation.

5.7 Parameter Analysis

In this section, to further evaluate the sensitivity of our proposed method to hyper-parameters, we conduct detailed experiments on MIND-Small dataset with different values of embedding dimension d, which is one of the most prominent parameters. In fact, it is crucial to evaluate the performance with lower embedding dimension when the computation and storage resources are limited. Specifically, we reduce the embedding dimension and extend the two groups of contrastive experiments in Section 5.6, HDNR versus EDNR and HNNR versus ENNR. The comparison results are displayed in Figures 3 and 4 respectively.

As shown in the figures, first of all, we observe that with the increases of d, the performance of our HDNR will also improve in terms of all evaluation metrics, though the trend of ascending becomes flat gradually. Then we discover that both HDNR and HNNR continuously outperform the EDNR and ENNR respectively with different values of embedding dimension, which proves the effectiveness of hyperbolic geometry in a fine-grained way. As a result, we demonstrate that our proposed HDNR is at a competitive advantage compared to the variants modeling in Euclidean space.

Moreover, it is worth noting that the gap between the variants HNNR and ENNR is much more significant at the beginning, but gradually narrows with the increases of embedding dimension according to Figure 4. Although hyperbolic space possesses more capacity over Euclidean space all the way, the inconsistent capacity problem in Euclidean space is much more intolerable while the embedding dimension is lower. Apparently, hyperbolic geometry possesses significant superiority when dealing with low-dimensional situations, due to its exponential growth property.

In summary, we demonstrate the effectiveness of our method and confirm the superiority of hyperbolic geometry in a fine-grained

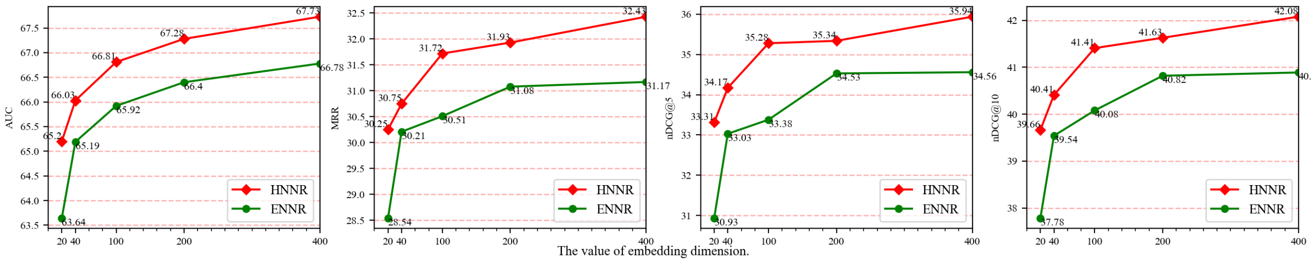

Figure 4: Performance with different embedding dimensions between HNNR and ENNR.

way. Even though with a low embedding dimension, our proposed HDNR is surprisingly able to produce high-quality representations, which makes it to be especially favourable in low-memory and low-storage scenarios.

6 CONCLUSION

In this paper, we propose a novel debiased approach based on hyperbolic space for personalized news recommendation, named HDNR, to explore users' intrinsic interests as well as alleviate conformity bias in data distribution. Specifically, first, we resort to hyperboloid manifold with exponential growth volume to obtain low distortion representations since the user-news interaction network exhibits power-law distribution. On this basis, we further design a re-weighting aggregation module to learn more accurate user representations through considering local relative importance of the clicked news among contextual history and its global popularity degree simultaneously, in order to achieve the above-mentioned two-fold purposes. Extensive experiments on real-world datasets validate the effectiveness of our approach. In the future, we tend to solve the potential problems when recommending popular news, which may be important in news recommendation task.

ACKNOWLEDGMENTS

We sincerely thank all the anonymous reviewers for their comments and suggestions. This work is supported by the National Natural Science Foundation of China (No.62106059), the National Key Research and Development Program of China (grant No.2021YFB3100600), the Strategic Priority Research Program of Chinese Academy of Sciences (grant No.XDC02040400), and the Youth Innovation Promotion Association of CAS (Grant No. 2021153).

REFERENCES

[1] Abhinandan S Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th international conference on World Wide Web*, pages 271–280, 2007.
[2] Dhruv Khattar, Vaibhav Kumar, Vasudeva Varma, and Manish Gupta. Weave&rec: A word embedding based 3-d convolutional network for news recommendation. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 1855–1858, 2018.
[3] Chong Wang and David M Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 448–456, 2011.
[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
[6] Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. Neural news recommendation with long-and short-term user representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 336–345, 2019.
[7] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. Neural news recommendation with attentive multi-view learning. *arXiv preprint arXiv:1907.05576*, 2019.
[8] Tao Qi, Fangzhao Wu, Chuhan Wu, Peiru Yang, Yang Yu, Xing Xie, and Yongfeng Huang. Hierec: Hierarchical user interest modeling for personalized news recommendation. *arXiv preprint arXiv:2106.04408*, 2021.
[9] Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. Neural news recommendation with multi-head self-attention. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 6389–6394, 2019.
[10] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. Npa: neural news recommendation with personalized attention. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2576–2584, 2019.
[11] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.
[12] Suyu Ge, Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. Graph enhanced representation learning for news recommendation. In *Proceedings of The Web Conference 2020*, pages 2863–2869, 2020.
[13] Linmei Hu, Chen Li, Chuan Shi, Cheng Yang, and Chao Shao. Graph neural news recommendation with long-term and short-term interest modeling. *Information Processing & Management*, 57(2):102142, 2020.
[14] Erzsébet Ravasz and Albert-László Barabási. Hierarchical organization in complex networks. *Physical review E*, 67(2):026112, 2003.
[15] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems*, 2020.
[16] Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima. Embedding-based news recommendation for millions of users. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1933–1942, 2017.
[17] Danyang Liu, Jianxun Lian, Shiyin Wang, Ying Qiao, Jiun-Hung Chen, Guangzhong Sun, and Xing Xie. Kred: Knowledge-aware document representation for news recommendations. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 200–209, 2020.
[18] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017.
[19] Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. *Advances in neural information processing systems*, 32, 2019.
[20] Caglar Gulcehre, Misha Denil, Mateusz Malinowski, Ali Razavi, Razvan Pascanu, Karl Moritz Hermann, Peter Battaglia, Victor Bapst, David Raposo, Adam Santoro, et al. Hyperbolic attention networks. *arXiv preprint arXiv:1805.09786*, 2018.
[21] Jianing Sun, Zhaoyue Cheng, Saba Zuberi, Felipe Pérez, and Maksims Volkovs. Hgcf: Hyperbolic graph convolution networks for collaborative filtering. In *Proceedings of the Web Conference 2021*, pages 593–601, 2021.
[22] Menglin Yang, Min Zhou, Jiahong Liu, Defu Lian, and Irwin King. Hrcf: Enhancing collaborative filtering via hyperbolic geometric regularization. In *Proceedings of the ACM Web Conference 2022*, pages 2462–2471, 2022.

[23] Menglin Yang, Zhihao Li, Min Zhou, Jiahong Liu, and Irwin King. Hicf: Hyperbolic informative collaborative filtering. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2212–2221, 2022.

[24] Yankai Chen, Menglin Yang, Yingxue Zhang, Mengchen Zhao, Ziqiao Meng, Jianye Hao, and Irwin King. Modeling scale-free graphs with hyperbolic geometry for knowledge-aware recommendation. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 94–102, 2022.

[25] Anchen Li, Bo Yang, Farookh Khadeer Hussain, and Huan Huo. Hsr: Hyperbolic social recommender. *Information Sciences*, 585:275–288, 2022.

[26] Hao Wang, Defu Lian, Hanghang Tong, Qi Liu, Zhenya Huang, and Enhong Chen. Hypersorec: Exploiting hyperbolic user and item representations with multiple aspects for social-aware recommendation. *ACM Transactions on Information Systems (TOIS)*, 40(2):1–28, 2021.

[27] Ellen M Voorhees et al. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82, 1999.

[28] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[29] Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. Dkn: Deep knowledge-aware network for news recommendation. In *Proceedings of the 2018 world wide web conference*, pages 1835–1844, 2018.

[30] Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. User-as-graph: User modeling with heterogeneous graph pooling for news recommendation. In *IJCAI*, pages 1624–1630, 2021.

[31] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.