Graph Neural News Recommendation with Unsupervised Preference Disentanglement

Linmei Hu[1], **Siyong Xu**[1], **Chen Li**[1], **Cheng Yang**[1], **Chuan Shi**[1]*
Nan Duan[2], **Xing Xie**[2], **Ming Zhou**[2]
[1]School of Computer Science,
Beijing University of Posts and Telecommunications, Beijing, China
[2]Microsoft Research Asia, Beijing, China
{hulinmei, xusiyong, leechen, yangcheng, shichuan}@bupt.edu.cn
{nanduan, xing.xie, mingzhou}@microsoft.com

Abstract

With the explosion of news information, personalized news recommendation has become very important for users to quickly find their interested contents. Most existing methods usually learn the representations of users and news from news contents for recommendation. However, they seldom consider high-order connectivity underlying the user-news interactions. Moreover, existing methods failed to disentangle a user's latent preference factors which cause her clicks on different news. In this paper, we model the user-news interactions as a bipartite graph and propose a novel **G**raph Neural **N**ews Recommendation model with **U**nsupervised Preference **D**isentanglement, named GNUD. Our model can encode high-order relationships into user and news representations by information propagation along the graph. Furthermore, the learned representations are disentangled with latent preference factors by a neighborhood routing algorithm, which can enhance expressiveness and interpretability. A preference regularizer is also designed to force each disentangled subspace to independently reflect an isolated preference, improving the quality of the disentangled representations. Experimental results on real-world news datasets demonstrate that our proposed model can effectively improve the performance of news recommendation and outperform state-of-the-art news recommendation methods.

1 Introduction

The amount of news and articles on many news platforms, such as Google News[1], has been growing constantly at an explosive rate, making it difficult for users to seek for news that they are interested in. In order to tackle the problem of information overload and meet the needs of users, news recommendation has been playing an increasingly important role for mining users' reading interest and providing personalized contents (IJntema et al., 2010; Liu et al., 2010).

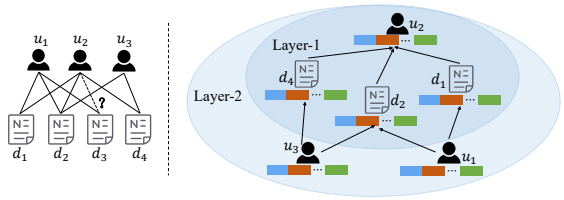


Figure 1: An illustration of user-news interaction graph and high-order connectivity. The representations of user and news are disentangled with latent preference factors.

A core problem in news recommendation is how to learn better representations of users and news. Recently, many deep learning based methods have been proposed to automatically learn informative user and news representations (Okura et al., 2017; Wang et al., 2018). For instance, DKN (Wang et al., 2018) learns knowledge-aware news representation via multi-channel CNN and gets a representation of a user by aggregating her clicked news history with different weights. However, these methods (Wu et al., 2019b; Zhu et al., 2019; An et al., 2019) usually focus on news contents, and seldom consider the collaborative signal in the form of high-order connectivity underlying the user-news interactions. Capturing high-order connectivity among users and news could deeply exploit structure characteristics and alleviate the sparsity, thus improving the rec-

*Corresponding author:Chuan Shi(shichuan@bupt.edu.cn)
[1]https://news.google.com/

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4255–4264

ommendation performance (Wang et al., 2019). For example, as shown in Figure 1, the high-order relationship u_1–d_1–u_2 indicates the behavior similarity between u_1 and u_2 so that we may recommend d_3 to u_2 since u_1 clicked d_3, while d_1–u_2–d_4 implies d_1 and d_4 may have similar target users.

Moreover, users may click different news due to their great diversity of preferences. The real-world user-news interactions arise from highly complex latent preference factors. For example, as shown in Figure 1, u_2 might click d_1 under her preference to entertainment news, while chooses d_4 due to her interest in politics. When aggregating neighborhood information along the graph, different importance of neighbors under different latent preference factors should be considered. Learning representations that uncover and disentangle these latent preference factors can bring enhanced expressiveness and interpretability, which nevertheless remains largely unexplored by the existing literatures on news recommendation.

In this work, to address the above issues, we model the user-news interactions as a bipartite graph and propose a novel **G**raph Neural **N**ews Recommendation Model with **U**nsupervised preference **D**isentanglement **(GNUD)**. Our model is able to capture the high-order connectivities underlying the user-news interactions by propagating the user and news representations along the graph. Furthermore, the learned representations are disentangled by a neighborhood routing mechanism, which dynamically identifies the latent preference factors that may have caused the click between a user and news, and accordingly assigning the news to a subspace that extracts and convolutes features specific to that factor. To force each disentangled subspace to independently reflect an isolated preference, a novel preference regularizer is also designed to maximize the mutual information measuring dependency between two random variables in information theory to strengthen the relationship between the preference factors and the disentangled embeddings. It further improves the disentangled representations of users and news. To summarize, this work makes the following three contributions:

(1) In this work, we model the user-news interactions as a bipartite graph and propose a novel graph neural news recommendation model GNUD with unsupervised preference disentanglement. Our model improves the recommendation performance by fully considering the high-order connectivities and latent preference factors underlying the user-news interactions.

(2) In our model GNUD, a preference regularizer is designed to enforce each disentangled embedding space to independently reflect an isolated preference, further improving the quality of disentangled representations for users and news.

(3) Experimental results on real-world datasets demonstrate that our proposed model significantly outperforms state-of-the-art news recommendation methods.

2 Related Work

In this section, we will review the related studies in three aspects, namely news recommendation, graph neural networks and disentangled representation learning.

News recommendation. Personalized news recommendation is an important task in natural language processing field, which has been widely explored in recent years. Learning better user and news representations is a central task for news recommendation. Traditional collaborative filtering (CF) based methods (Wang and Blei, 2011) often utilize historical interactions between users and news to define the objective function for model training, aiming to predict a personalized ranking over a set of candidates for each user. They usually suffer from cold-start problem since news are often substituted frequently. Many works attempt to take advantage of rich content information, effectively improving the recommendation performance. For example, DSSM (Huang et al., 2013) is a content-based deep neural network to rank a set of documents given a query. Some works (Wang et al., 2018; Zhu et al., 2019) propose to improve news representations via external knowledge, and learn representations of users from their browsed news using an attention module. Wu et al. (2019b) applied attention mechanism at both word- and news-level to model different informativeness on news content for different users. Wu et al. (2019a) exploited different types of news information with an attentive multi-view learning framework. An et al. (2019) considered both titles and topic categories of news, and learned both long- and short-term user representations, while Wu et al. (2019c) represented them by multi-head attention mechanism. However, these works seldom mine high-order structure information.

Graph neural networks. Recently, graph neu-

ral networks (GNN) (Kipf and Welling, 2016; Hamilton et al., 2017; Veličković et al., 2017) have received growing attentions in graph embedding because of its powerful representation learning based on node features and graph structure. Wang et al. (2019) explored the GNN to capture high-order connectivity information in user-item graph by propagating embeddings on it, which achieves better performance on recommendation. However, existing news recommendation methods focus on, and rely heavily on news contents. Few news recommendation models consider the user-news interaction graph structure which encodes useful high-order connectivity information. Hu et al. (2020) modeled the user-news interactions as a graph and proposed a graph convolution based model combining long-term and short-term interests, which demonstrates the effectiveness of exploiting the user-news interaction graph structure. Different from all these methods, in this work, we consider both the high-order connectivity information and latent preference factor underlying the user-news interactions. We propose a novel graph neural news recommendation model with unsupervised preference disentanglement.

Disentangled representation learning. Disentangled representation learning aims to identify and disentangle different latent explanatory factors hidden in the observed data (Bengio et al., 2013), which has been successfully applied in the field of computer vision (Kim and Mnih, 2018; Gidaris et al., 2018; Hsieh et al., 2018). $\beta-$VAE (Higgins et al., 2017) is a deep unsupervised generative approach that can automatically discover the independent latent factors of variation in unsupervised data, which is based on the VAE framework (Kingma and Welling, 2013). Recently, disentangled representation learning has been investigated on graph-structured data (Ma et al., 2019a,b). To the best of our knowledge, this is the first work to explore disentanglement in news recommendation.

3 Problem Formulation

The news recommendation problem can be formalized as follows. Given the user-news historical interactions $\{(u, d)\}$, we aim to predict whether a user u_i will click a candidate news d_j that she has not seen before.

In this paper, for a news article d, we consider the title T and profile P (a given set of entities E and their corresponding entity types C from the news content) as features. The entities E and their corresponding entity types C are already given in the datasets. Each news title T consists of a word sequence $T = \{w_1, w_2, \cdots, w_m\}$. Each profile P contains a sequence of entities defined as $E = \{e_1, e_2, \cdots, e_p\}$ and corresponding entity types $C = \{c_1, c_2, \cdots, c_p\}$. We denote the title embedding as $\mathbf{T} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_m]^T \in R^{m \times n_1}$, entity set embedding as $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_p]^T \in R^{p \times n_1}$, and the entity-type set embedding as $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_p]^T \in R^{p \times n_2}$. $\mathbf{w}$, $\mathbf{e}$ and $\mathbf{c}$ are respectively the embedding vectors of word w, entity e, and entity type c. n_1 and n_2 are the dimension of word (entity) and entity-type embeddings. These embeddings can be pre-trained from a large corpus or randomly initialized. Following (Zhu et al., 2019), we define the profile embedding $\mathbf{P} = [\mathbf{e}_1, g(\mathbf{c}_1), \mathbf{e}_2, g(\mathbf{c}_2), \cdots, \mathbf{e}_p, g(\mathbf{c}_p)]^T$ where $\mathbf{P} \in R^{2p \times n_1}$. $g(\mathbf{c})$ is the transformation function as $g(\mathbf{c}) = \mathbf{M}_c\mathbf{c}$, where $\mathbf{M}_c \in R^{n_1 \times n_2}$ is a trainable transformation matrix.

4 Our Proposed Method

In this section, we first introduce the news content information extractor which learns a news representation $\mathbf{h}_d$ from news content. Then we detail our proposed graph neural model GNUD with unsupervised preference disentanglement for news recommendation. Our model not only exploits the high-order structure information underlying the user-news interaction graph but also considers the different latent preference factors causing the clicks between users and news. A novel preference regularizer is also introduced to force each disentangled subspace independently reflect an isolated preference factor.

4.1 News Content Information Extractor

We first describe how to obtain a news representation $\mathbf{h}_d$ from news content including news title T and profile P. The content-based news representations would be taken as initial input embeddings of our model GNUD. Following DAN (Zhu et al., 2019), we use two parallel convolutional neural networks (PCNN) taking the title $\mathbf{T}$ and profile $\mathbf{P}$ of news as input to learn the title-level and profile-level representation $\widehat{\mathbf{T}}$ and $\widehat{\mathbf{P}}$ for news. Finally we concatenate $\widehat{\mathbf{T}}$ and $\widehat{\mathbf{P}}$, and get the final news representation $\mathbf{h}_d$ through a fully connected layer f:

$$\mathbf{h}_d = f([\widehat{\mathbf{T}}; \widehat{\mathbf{P}}]). \quad (1)$$

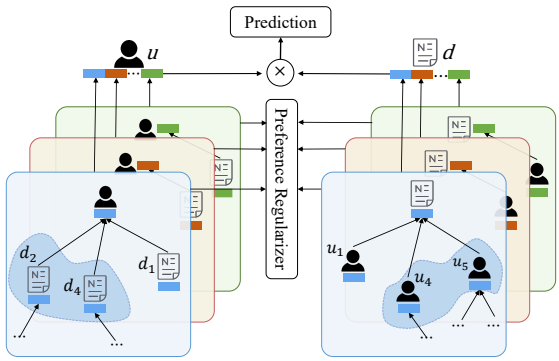


Figure 2: Illustration of our proposed model GNUD.

4.2 GNUD

To capture the high-order connectivity underlying the user-news interactions, we model the user-news interactions as a bipartite graph $\mathcal{G} = \{\mathcal{U}, \mathcal{D}, \mathcal{E}\}$, where $\mathcal{U}$ and $\mathcal{D}$ are the sets of users and news, $\mathcal{E}$ is the set of edges and each edge $e = (u, d) \in \mathcal{E}$ indicates that user u explicitly clicks news d. Our model GNUD enables information propagation among users and news along the graph, thus capturing the high-order relationships among users and news. Additionally, GNUD learns disentangled embeddings that uncover the latent preference factors behind user-news interactions, enhancing expressiveness and interpretability. In the following, we present one single graph covolution layer with preference disentanglement.

4.2.1 Graph Convolution Layer with Preference Disentanglement

Given the user-news bipartite graph $\mathcal{G}$ where the user embedding $\mathbf{h}_u$ is randomly initialized and news embedding $\mathbf{h}_d$ is obtained with the news content information extractor (Section 4.1), a graph convolutional layer aims to learn the representation $\mathbf{y}_u$ of a node u by aggregating its neighbors' features:

$$\mathbf{y}_u = \text{Conv}(\mathbf{h}_u, \{\mathbf{h}_d : (u, d) \in \mathcal{E}\}). \quad (2)$$

Considering that users' click behaviors could be caused by different latent preference factors, we propose to derive a layer $\text{Conv}(\cdot)$ such that the output $\mathbf{y}_u$ and $\mathbf{y}_d$ are disentangled representations. Each disentangled component reflect one preference factor related to the user or news. The learned disentangled user and news embeddings can bring enhanced expressiveness and interpretability. Assuming that there are K factors, we would like to let $\mathbf{y}_u$ and $\mathbf{y}_d$ be composed of K independent components: $\mathbf{y}_u = [\mathbf{z}_{u,1}, \mathbf{z}_{u,2}, \cdots, \mathbf{z}_{u,K}]$, $\mathbf{y}_d = [\mathbf{z}_{d,1}, \mathbf{z}_{d,2}, \cdots, \mathbf{z}_{d,K}]$, where $\mathbf{z}_{u,k}$ and $\mathbf{z}_{d,k} \in R^{\frac{l_{out}}{K}}$ $(1 \leq k \leq K)$ (l_{out} is the dimension of $\mathbf{y}_u$ and $\mathbf{y}_d$), respectively characterizing the k-th aspect of user u and news d related to the k-th preference factor. Note that in the following of this paper, we focus on user u and describe the learning process of its representation $\mathbf{y}_u$. The news d can be learned similarly, which is omitted.

Formally, given a u-related node $i \in \{u\} \bigcup \{d : (u, d) \in \mathcal{E}\}$, we use a subspace-specific projection matrix W_k to map the feature vector $\mathbf{h}_i \in R^{l_{in}}$ into the k-th preference related subspace:

$$\mathbf{s}_{i,k} = \frac{\text{ReLU}(\mathbf{W}_k^\top \mathbf{h}_i + \mathbf{b}_k)}{\| \text{ReLU}(\mathbf{W}_k^\top \mathbf{h}_i + \mathbf{b}_k) \|_2}, \quad (3)$$

where $\mathbf{W}_k \in R^{l_{in} \times \frac{l_{out}}{K}}$, and $\mathbf{b}_k \in R^{\frac{l_{out}}{K}}$. Note that $\mathbf{s}_{u,k}$ is not equal to the final representation of the k-th component of u: $\mathbf{z}_{u,k}$, since it has not mined any information from neighboring news yet. To construct $\mathbf{z}_{u,k}$, we need to mine the information from both $\mathbf{s}_{u,k}$ and the neighborhood features $\{\mathbf{s}_{d,k} : (u, d) \in \mathcal{E}\}$.

The main intuition is that when constructing $\mathbf{z}_{u,k}$ characterizing the k-th aspect of u, we should only use the neighboring news articles d which connect with user u due to the preference factor k instead of all the neighbors. In this work, we apply a neighborhood routing algorithm (Ma et al., 2019a) to identify the subset of neighboring news that actually connect to u due to the preference factor k.

Neighborhood routing algorithm. The neighborhood routing algorithm infers the latent preference factors behind user-news interactions by iteratively analyzing the potential subspace formed by a user and her clicked news. The detail is illustrated in Algorithm 1. Formally, let $r_{d,k}$ be the probability that the user u clicks the news d due to the factor k. Then it's also the probability that we should use the news d to construct $\mathbf{z}_{u,k}$. $r_{d,k}$ is an unobserved latent variable which can be inferred in an iterative process. The motivation of the iterative process is as follows. Given $\mathbf{z}_{u,k}$, the value of the latent variables $\{r_{d,k} : 1 \leq k \leq K, (u, d) \in \mathcal{E}\}$ can be obtained by measuring the similarity between user u and her clicked news d under the k-th subspace, which is computed as Eq. 4. Initially, we set $\mathbf{z}_{u,k} = \mathbf{s}_{u,k}$. On the other hand, after obtaining the latent variables $\{r_{d,k}\}$, we can find an estimate of $\mathbf{z}_{u,k}$ by aggregating information from the clicked news, which is computed as Eq. 5:

Algorithm 1 Neighborhood Routing Algorithm

Require:
$\mathbf{s}_{i,k}, i \in \{u\} \bigcup \{d : (u,d) \in \mathcal{E}\}, 1 \leq k \leq K;$
Ensure:
$\mathbf{z}_{u,k}, 1 \leq k \leq K;$
$\forall k = 1, ...K, \mathbf{z}_{u,k} \leftarrow \mathbf{s}_{u,k}$
for T iterations **do**
  for d that satisfies $(u,d) \in \mathcal{E}$ **do**
    $\forall k = 1, \cdots, K : r_{d,k} \leftarrow \mathbf{z}_{u,k}^{\top} \mathbf{s}_{d,k}$
    $\forall k = 1, \cdots, K : r_{d,k} \leftarrow \text{softmax}(r_{d,k})$
  end for
  for factor $k = 1, 2, ...K$ **do**
    $\mathbf{z}_{u,k} \leftarrow \mathbf{s}_{u,k} + \sum_{d:(u,d)\in\mathcal{E}} r_{d,k} \mathbf{s}_{d,k}$
    $\mathbf{z}_{u,k} \leftarrow \mathbf{z}_{u,k} / \parallel \mathbf{z}_{u,k} \parallel_2$
  end for
end for
return $\mathbf{z}_{u,k}$

$$r_{d,k}^{(t)} = \frac{\exp({\mathbf{z}_{u,k}}^{(t)\top} \mathbf{s}_{d,k})}{\sum_{k'=1}^{K} \exp({\mathbf{z}_{u,k}}^{(t)\top} \mathbf{s}_{d,k})}, \tag{4}$$

$$\mathbf{z}_{u,k}^{(t+1)} = \frac{\mathbf{s}_{u,k} + \sum_{d:(u,d)\in\mathcal{G}} r_{d,k}^{(t)} \mathbf{s}_{d,k}}{\parallel \mathbf{s}_{u,k} + \sum_{d:(u,d)\in\mathcal{G}} r_{d,k}^{(t)} \mathbf{s}_{d,k} \parallel_2}, \tag{5}$$

where iteration $t = 0, \cdots, T-1$. After T iterations, the output $\mathbf{z}_{u,k}^{(T)}$ is the final embedding of user u in the k-th latent subspace and we obtain $\mathbf{y}_u = [\mathbf{z}_{u,1}, \mathbf{z}_{u,2}, \cdots, \mathbf{z}_{u,K}]$.

The above shows a single graph convolutional layer with preference disentanglement, which aggregates information from the first-order neighbors. In order to capture information from high-order neighborhood and learn high-level features, we stack multiple layers. Specially, we use L layers and get the final disentangled representation $\mathbf{y}_u^{(L)} \in R^{K\Delta n}$ ($K\Delta n = l_{out}$) for user u and $\mathbf{y}_d^{(L)}$ for news d, where Δn is the dimension of a disentangled subspace.

4.2.2 Preference Regularizer

Naturally, we hope each disentangled subspace can reflect an isolated latent preference factor independently. Since there are no explicit labels indicating the user preferences in the training data, a novel preference regularizer is also designed to maximize the mutual information measuring dependency between two random variables in information theory to strengthen the relationship between the preference factors and the disentangled embeddings. According to (Yang et al., 2018), the mutual information maximization can be converted to the following form.

Given the representation of a user u in k-th $(1 \leq k \leq K)$ latent subspace, the preference regularizer $P(k|\mathbf{z}_{u,k})$ estimates the probability of the k-th subspace (w.r.t. the k-th preference) that $\mathbf{z}_{u,k}$ belongs to:

$$P(k|\mathbf{z}_{u,k}) = \text{softmax}(\mathbf{W}_p \cdot \mathbf{z}_{u,k} + \mathbf{b}_p), \tag{6}$$

where $\mathbf{W}_p \in R^{K\times\Delta n}$, and parameters in the regularizer $P(\cdot)$ are shared with all the users and news.

4.3 Model Training

Finally, we add a fully-connected layer, i.e., $\mathbf{y}'_u = {\mathbf{W}^{(L+1)}}^{\top} \mathbf{y}_u^{(L)} + \mathbf{b}^{(L+1)}$, where $\mathbf{W}^{(L+1)} \in R^{K\Delta n \times K\Delta n}$, $\mathbf{b}^{(L+1)} \in R^{K\Delta n}$. We use the simple dot product to compute the news click probability score, which is computed as $\hat{s}\langle u,d\rangle = {\mathbf{y}'_u}^{\top} \mathbf{y}'_d$.

Once obtaining the click probability scores $\hat{s}\langle u,d\rangle$, we define the following base loss function for training sample (u,d) with the ground truth $y_{u,d}$:

$$\mathcal{L}_1 = -[y_{u,d} \ln(\hat{y}_{u,d}) + (1 - y_{u,d}) \ln(1 - \hat{y}_{u,d})], \tag{7}$$

where $\hat{y}_{u,d} = \sigma(\hat{s}\langle u,d\rangle)$.

Then we add the preference regularization term of both u and d, which can be written as:

$$\mathcal{L}_2 = -\frac{1}{K} \sum_{k=1}^{K} \sum_{i\in\{u,d\}} \ln P(k|\mathbf{z}_{i,k})[k]. \tag{8}$$

The overall training loss can be rewritten as:

$$\mathcal{L} = \sum_{(u,d)\in\mathcal{T}_{\text{train}}} ((1-\lambda)\mathcal{L}_1 + \lambda\mathcal{L}_2) + \eta\|\Theta\|, \tag{9}$$

where $\mathcal{T}_{\text{train}}$ is training set. For each positive sample (u,d), we sample a negative sample from unobserved reading history of u for training. λ is a balance coefficient. η is the regularization coefficient and $\|\Theta\|$ denotes all the trainable parameters.

Note that during training and testing, the news that have not been read by any users are taken as isolated nodes in the graph. Their representations are based on only content feature h_d without neighbor aggregation, and can also be disentangled via Eq. 3.

5 Experiments

5.1 Datasets and Experimental Settings

Datasets. We conduct experiments on the real-world online news datasets Adressa (Gulla et al., 2017)[2] from a Norwegian news portal to evaluate our model. We use two datasets named $Adressa$-$1week$ and $Adressa$-$10week$, which respectively collect news click logs as long as 1 week and 10 weeks. Following DAN (Zhu et al., 2019), we just select user id, news id, time-stamp, the title and profile of news to build our datasets, and preprocess the data by removing the stopwords in the news content. The statistics of our final datasets are shown in Table 1.

For the $Adressa$-$1week$ dataset, we use the first 5 days' historical data for the construction of user-news bipartite graph. The 6-th day's is used to build training samples: $\{(u, d)\}$. 20% randomly sampled from the last day's are for validation and the remaining are regarded as test set. Note that during testing, we reconstruct the graph with all the previous 6 days' historical data. Similarly, for the $Adressa$-$10week$ dataset, we construct the graph with the first 50 days' data, the following 10 days are served to generate training pairs, 20% of the last 10 days' for validation data and 80% for test. Note that, for the baselines, we also use the data from the first 5 (50) days for constructing user's historical data, the following 1 (10) days is used to generate training pairs. The validation and test set constructed with the last 1 (10) days are also the same for all the models.

Experimental settings. In our experiments, the dimension of word/entity embeddings and entity type embeddings is set as $n_1 = n_2 = 50$, and the dimension of input user and news embeddings l_{in} is set as 128. The embeddings of words, entities, entity types and users are randomly initialized with a Gaussian distribution $\mathcal{N}(0, 0.1)$. In our methods, due to the large scale of the datasets, we sample a fixed-size set of neighbors ($size = 10$) for a user, and we set $size = 30$ for a news, according to the average degree of users and news respectively. The number of latent preference factors is $K = 7$, and the dimension of each disentangled subspace is $\Delta n = 16$. The number of graph convolution layers is set to 2. The dropout rate is 0.5. The balance coefficient λ is set as 0.004. We test our model with different value of λ ranging from 0.001 to 0.02 (with step 0.001) and find that our model is insensitive to λ in $[0.001, 0.02]$. Finally, Adam (Kingma and Ba, 2014) is applied for model optimization, and the learning rate is 0.0005. The batch size is set to 128. These hyper-parameters were all selected according to the results on validation set.

[2]http://reclab.idi.ntnu.no/dataset/

Number	$*$ 1week	$*$ 10week
# users	537,629	590,674
# news	14,732	49,994
# clicks	2,107,312	15,127,204
# vocabulary	116,603	279,214
# entity-type	11	11
# average words	4.03	4.10
# average entities	22.11	21.29

Table 1: Statistics of our datasets.

It is worth noting that our model can deal with new coming news documents that have not previously existed in the user-news interaction graph $\mathcal{G}$ during training or testing. Our model takes these news documents as isolated nodes in the graph $\mathcal{G}$. Their representations are based on only content feature h_d without neighbor aggregation, and can also be disentangled via Eq. 3.

5.2 Performance Evaluation

We evaluate the performance of our model GNUD by comparing it with the following state-of-the-art baseline methods:

LibFM (Rendle, 2012), a feature-based matrix factorization method, with the concatenation of TF-IDF vectors of news title and profile as input.

CNN (Kim, 2014), applying two parallel CNNs to word sequences in news titles and profiles respectively and concatenate them as news features. The user representation is learned from the user's news history.

DSSM (Huang et al., 2013), a deep structured semantic model. In our experiments, we model the user's clicked news as the query and the candidate news as the documents.

Wide & Deep (Cheng et al., 2016), a deep model for recommendation which combines a (Wide) linear model and (Deep) feed-forward neural network. We also use the concatenation of news title and profile embeddings as features.

DeepFM (Guo et al., 2017), a general model that combines factorization machines and deep neural networks that share the input. We use the same input as Wide & Deep for DeepFM.

Methods	Adressa-1week		Adressa-10week	
	AUC	F1	AUC	F1
LibFM	61.20±1.29	59.87±0.98	63.76±1.05	62.41±0.72
CNN	67.59±0.94	66.33±1.44	69.07±0.95	67.78±0.69
DSSM	68.61±1.02	69.92±1.13	70.11±1.35	70.96±1.56
Wide&Deep	68.25±1.12	69.32±1.28	73.28±1.26	69.52±0.83
DeepFM	69.09±1.45	61.48±1,31	74.04±1.69	65.82±1.18
DMF	55.66±0.84	56.46±0.97	53.20±0.89	54.15±0.47
DKN	75.57±1.13	76.11±0.74	74.32±0.94	72.29±0.41
DAN	75.93±1.25	74.01±0.83	76.76±1.06	71.65±0.57
GNewsRec	81.16±1.19	82.85±1.15	78.62±1.38	81.01±0.64
GNUD w/o Disen	78.33±1.29	79.09±1.22	78.24±0.13	80.58±0.45
GNUD w/o PR	83.12±1.53	81.67±1.56	80.61±1.07	80.92±0.31
GNUD	**84.01±1.16**	**83.90±0.58**	**83.21±1.91**	**81.09±0.23**

Table 2: The performance of different methods on news recommendation.

DMF (Xue et al., 2017), a CF based deep matrix factorization model without considering the news content.

DKN (Wang et al., 2018), a deep content based news recommendation framework fusing semantic-level and knowledge-level representations. We model the news title and profile as semantic-level and knowledge-level representations, respectively.

DAN (Zhu et al., 2019), a deep attention neural network for news recommendation which can capture the dynamic diversity of news and user's interests, and consider the users' click sequence information.

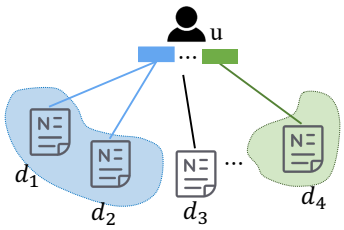
GNewsRec (Hu et al., 2020), a graph neural network based method combining long-term and short term interest modeling for news recommendation.

All the baselines are initialized as the corresponding papers, and in terms of neural network models we use the same word embedding dimension for fair comparison. Then they are carefully tuned to achieve their optimal performance. We independently repeat each experiment for 10 times and report the average performance.

Result analysis. The comparisons between different methods are summarized in Table 2. We can observe that our proposed model GNUD consistently outperforms all the state-of-the-art baseline methods on both datasets. GNUD improves the best deep neural models DKN and DAN more than 6.45% on AUC and 7.79% on F1 on both datasets. The main reason is that our model fully exploits the high-order structure information in the user-news interaction graph, learning better representations of users and news. Compared to the best-performed baseline method GNewsRec, our model GNUD achieves better performance on both datasets in terms of both AUC (+2.85% and +4.59% on the two datasets, respectively) and F1 (+1.05% and +0.08%, respectively). This is because that our model considers the latent preference factors that cause the user-news interactions and learns representations that uncover and disentangle these latent preference factors, which enhance expressiveness.

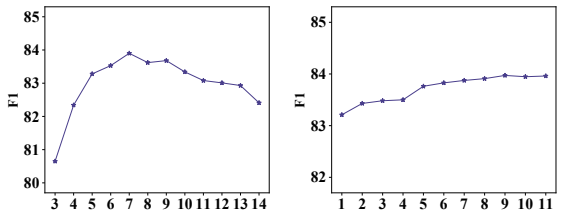
From Table 2, we can also see that all the content-based methods outperform the CF based model DMF. This is because CF based methods suffer a lot from cold-start problem since most news are new coming. Except for DMF, all the deep neural network based baselines (e.g., CNN, DSSM Wide&Deep, DeepFM, etc.) significantly outperform LibFM, which shows that deep neural models can capture more implicit but informative features for user and news representations. DKN and DAN further improve other deep neural models by incorporating external knowledge and applying a dynamic attention mechanism.

Comparison of GNUD variants. To further demonstrate the efficacy of the design of our model GNUD, we compare among the variants of our model. As we can see from the last three lines in Table 2, when the preference disentanglement is removed, the performance of the model GNUD w/o Disen (GNUD without preference disentanglement) drops largely by 5.68% and 4.97% in terms of AUC on the two datasets (4.81% and 0.51% on F1), respectively. This observation demonstrates the effectiveness and necessity of preference disentangled representations of users and news. Com-



News	Keywords
d_1	norway oljebransjen (Norway oil industry), norskehavet (Norwegian sea), helgelandskysten (Helgeland coast), hygen (hygen), energy (energy), trondheim (a city)
d_2	Statkraft (State Power Corporation of Norway), trønderenergi (tronder energy), snillfjord (snill fjord), trondheimsfjorden (trondheim fjord), vindkraft (wind power), energy (energy)
d_3	Bolig (residence), hage (garden), hjemme (home), fossen (waterfall), hus (house), home (home)
d_4	health-and-fitness (health and fitness), mørk sjokolade (dark chocolate), vitaminrike (vitamin), olivenolje (olive oil), grønnsaker (vegetables), helse (health)

Figure 3: Visualization of a user's clicked news which belong to different disentangled subspaces w.r.t. different preference factors. We use six keywords (translated into English) to illustrate a news.



(a) # Preference factors. (b) # Routing Iterations.

Figure 4: Influence of different number of preference factors and routing iterations.

Methods	Adressa-1week		Adressa-10week	
	AUC	F1	AUC	F1
GNUD-1	80.96	79.86	82.22	80.61
GNUD-2	84.01	**83.90**	**83.21**	**81.09**
GNUD-3	**84.03**	82.18	83.05	80.93

Table 3: The performance of GNUD with different layer numbers.

pared to GNUD w/o PR (GNUD without preference regularizer), we can see that introducing the preference regularizer which enforces each disentangled embedding subspace independently reflect an isolated preference, can bring performance gains on both AUC (+0.89% and +2.6%, respectively) and F1 (+2.23% and +0.17%, respectively).

5.3 Case Study

To intuitively demonstrate the efficacy of our model, we randomly sample a user u and extract her logs from the test set. The representation of user u is disentangled into $K = 7$ subspaces and we randomly sample 2 subspaces. For each one, we visualize the top news that user u pay most attention to (with the probability $r_{d,k}$ larger than a threshold). As shown in Figure 3, different subspaces relect different preference factors. For example, one subspace (shown in blue) is related to "energy" as the top two news contain the keywords such as "oil industry", "hygen" and "wind power". The other subspace (shown in green) may indicate the latent preference factor about "healthy diet" as the related news contain the keywords such as "health", "vitamin" and "vegetables". The news d_3 about home has low probability in the both subspaces. It does not belong to any of the two preferences.

5.4 Parameter Analysis

In this section, we examine how different choices of some hyper-parameters affect the performance of GNUD.

Analysis of layer numbers. We investigate whether GNUD can benefit from multiple embedding propagation layers. We vary the layer numbers in the range of $\{1, 2, 3\}$ on both datasets. As we can see in Table 3, GNUD-2 (2 layers) is superior to others. The reason is that GNUD-1 considers the first-order neighbors only, while using over 2 layers may lead to overfitting, which indicates that applying a too deep architecture might bring noise to the representations in news recommendation task. Therefore, GNUD-2 is regarded as the most suitable choice.

Number of latent preference factors. We fix the dimension of each latent preference subspace as 16 and check the impact of the number K of latent preference factors. As shown in Figure 4 (a), we can find that with the increase of K, the performance first grows, reaching the best at K=7, and then begins to drop. Thus we set K=7 in our experiments.

Number of routing iterations. We study the performance with different number of routing iterations. As shown in Figure 4 (b), we can see that our model generally gets better performance with more routing iterations and finally achieves convergence after 7 iterations.

6 Conclusion

In this paper, we consider the high-order connectivity as well as the latent preference factors underlying the user-news interactions, and propose a novel graph neural news recommendation model GNUD with unsupervised preference disentanglement. Our model regards the user-news interactions as a bipartite graph and encode high-order relationships among users and news by graph convolution. Furthermore, the learned representations are disentangled with different latent preference factors by a neighborhood routing mechanism, enhancing expressiveness and interpretability. A preference regularizer is also designed to force each disentangled subspace to independently reflect an isolated preference, further improving the quality of user and news embeddings. Experimental results on real-world news datasets demonstrate that our model achieves significant performance gains compared to state-of-the-art methods, supporting the importance of exploiting the high-order connectivity and disentangling the latent preference factors in user and news representations.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. U1936220, 61806020, 61772082, 61972047, 61702296), the National Key Research and Development Program of China (2018YFB1402600), the CCF-Tencent Open Fund, and the Fundamental Research Funds for the Central Universities. We also acknowledge the valuable comments from Jianxun Lian at Microsoft Research Asia.

References

Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. 2019. Neural news recommendation with long-and short-term user representations. In *ACL*, pages 336–345.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In *DLRS*, pages 7–10.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. 2018. Unsupervised representation learning by predicting image rotations. In *ICLR*.

Jon Atle Gulla, Lemei Zhang, Peng Liu, Özlem Özgöbek, and Xiaomeng Su. 2017. The adressa dataset for news recommendation. In *WI*, pages 1042–1048.

Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. Deepfm: a factorization-machine based neural network for ctr prediction. In *IJCAI*, pages 1725–1731.

Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *NIPS*, pages 1024–1034.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*.

Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li F Fei-Fei, and Juan Carlos Niebles. 2018. Learning to decompose and disentangle representations for video prediction. In *NIPS*, pages 517–526.

Linmei Hu, Chen Li, Chuan Shi, Cheng Yang, and Chao Shao. 2020. Graph neural news recommendation with long-term and short-term interest modeling. *Information Processing Management*, 57(2):102142.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *CIKM*, pages 2333–2338.

Wouter IJntema, Frank Goossen, Flavius Frasincar, and Frederik Hogenboom. 2010. Ontology-based news recommendation. In *EDBT/ICDT Workshops*, page 16.

Hyunjik Kim and Andriy Mnih. 2018. Disentangling by factorising. In *ICML*, pages 2654–2663.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*, pages 1746–1751.

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

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. In *ICLR*.

Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. 2010. Personalized news recommendation based on click behavior. In *IUI*, pages 31–40.

Jianxin Ma, Peng Cui, Kun Kuang, Xin Wang, and Wenwu Zhu. 2019a. Disentangled graph convolutional networks. In *ICML*, pages 4212–4221.

Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. 2019b. Learning disentangled representations for recommendation. In *NIPS*, pages 5712–5723.

Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima. 2017. Embedding-based news recommendation for millions of users. In *KDD*, pages 1933–1942.

Steffen Rendle. 2012. Factorization machines with libfm. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3):57.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. In *ICLR*.

Chong Wang and David M Blei. 2011. Collaborative topic modeling for recommending scientific articles. In *KDD*, pages 448–456.

Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. Dkn: Deep knowledge-aware network for news recommendation. In *WWW*, pages 1835–1844.

Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural graph collaborative filtering. In *SIGIR*, pages 165–174.

Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019a. Neural news recommendation with attentive multi-view learning. In *IJCAI*, pages 3863–3869.

Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019b. Npa: Neural news recommendation with personalized attention. In *KDD*, pages 2576–2584.

Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019c. Neural news recommendation with multi-head self-attention. In *EMNLP-IJCNLP*, pages 6390–6395.

Hong-Jian Xue, Xinyu Dai, Jianbing Zhang, Shujian Huang, and Jiajun Chen. 2017. Deep matrix factorization models for recommender systems. In *IJCAI*, pages 3203–3209.

Cheng Yang, Maosong Sun, Xiaoyuan Yi, and Wenhao Li. 2018. Stylistic chinese poetry generation via unsupervised style disentanglement. In *EMNLP*, pages 3960–3969.

Qiannan Zhu, Xiaofei Zhou, Zeliang Song, Jianlong Tan, and Li Guo. 2019. Dan: Deep attention neural network for news recommendation. In *AAAI*, volume 33, pages 5973–5980.