Personalized News Recommendation with Candidate-aware User Modeling

Anonymous ACL Rolling

Abstract

News recommendation aims to match news with personalized user interest. Existing methods for news recommendation usually model user interest from historical clicked news without the consideration of candidate news. However, each user usually has multiple interests, and it is difficult for these methods to accurately match a candidate news with a specific user interest. In this paper, we present a candidate-aware user modeling method for personalized news recommendation, which can incorporate candidate news into user modeling for better matching between candidate news and user interest. More specifically, we propose a candidate-aware self-attention network that uses candidate news as guidance to model candidate-aware global user interest. In addition, we propose a candidate-aware CNN network to incorporate candidate news into local behavior context modeling to learn candidate-aware short-term user interest. Besides, we use a candidate-aware attention network to aggregate previously clicked news weighted by their relevance with candidate news to build candidate-aware user representation. The experiments on real-world datasets show the effectiveness of our approach in improving news recommendation performance.

1 Introduction

Personalized news recommendation is a critical technique for online news platforms to improve user experience (Lin et al., 2014; Garcia Esparza et al., 2010; Wu et al., 2020a; Zheng et al., 2018; Wu et al., 2020b; Khattar et al., 2018b). Accurate modeling of user interest on candidate news is important for personalized news recommendation (Ge et al., 2020; Hu et al., 2020b; Santosh et al., 2020; Wu et al., 2020c; Lee et al., 2020). Many existing methods first model user interests and candidate news content separately and then use their representations for interest matching (Wu et al., 2019d,e; Qi et al., 2020). For example, An et al. (2019) used

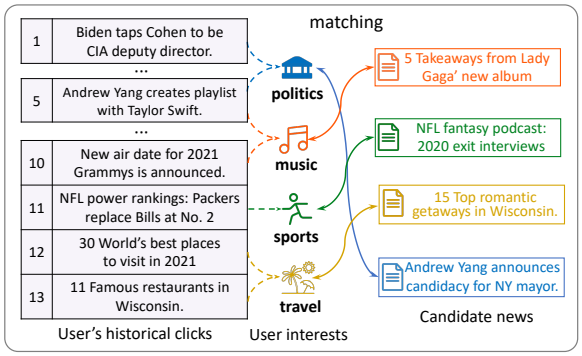

Figure 1: The matching of candidate news and user interest modeled from previously clicked news.

CNN network for news content modeling and used a GRU network and ID embeddings to learn user interest representations from clicked news. They modeled the relevance between user interests and candidate news based on the dot product of their representations. Wu et al. (2019a) applied attention networks to learn news and user interest representations. They also performed dot product between user interest and candidate news representations to model their relevance. In these methods, user interests are modeled in a candidate-agnostic way. However, each user usually has multiple interests (Liu et al., 2020b), and it may be difficult to accurately match candidate news with a specific user interest if candidate news is not considered in user modeling (Wang et al., 2018).

Our work is motivated by the following observations. First, users usually have multiple interests. For instance, as shown in Fig. 1, we can infer that the example user is interested in many different fields, such as politics, music, sports, and travel, from the news clicked by this user. However, a candidate news usually only matches a small part of user interests. For instance, the forth candidate news only matches user interests in politics, and it has low relevance to other interests of this user like music and sports. Thus, it may be difficult to accurately match the candidate news if candidate

news information is not considered in user modeling. Second, local contexts of users' news click behaviors are useful for inferring short-term user interests. For example, as shown in Fig. 1, we can infer the user's recent interests on travel in Wisconsin from the relatedness between the 12th and 13th news clicks. Third, long-range relatedness between users' historical clicks also provides rich information to model long-term user interests. For example, we can infer the long-term user interests in music from the long-range relatedness between the 5th and 10th clicks. Thus, understanding both short- and long-term user interests is important for accurate news recommendation (An et al., 2019).

In this paper, we propose a candidate-aware user modeling framework for personalized news recommendation (*CAUM*), which can incorporate candidate news information into user modeling for accurate interest matching. We propose a candidate-aware self-attention network to learn candidate-aware global user interest representations. It uses candidate news representation to guide the modeling of global relatedness between historical clicked news. In addition, we propose a candidate-aware CNN network to learn candidate-aware short-term user interest representations. It incorporates candidate news information into the modeling of local contexts of click behaviors. Besides, we adopt a candidate-aware attention network to weight clicked news based on their relevance with candidate news to learn candidate-aware user interest representation for better matching with candidate news. Experimental results on two real-world datasets verify that *CAUM* can improve the performance of user modeling for news recommendation.

The contribution of our paper is three-fold:

(1) We propose a candidate-aware user modeling method for personalized news recommendation, which can incorporate candidate news into user modeling for accurate user and news matching.

(2) We propose a candidate-aware self-attention network and a candidate-aware CNN network to infer global and short-term user interests from clicked news with the guidance of candidate news.

(3) Extensive experiments on two real-world datasets validate the effectiveness of our method.

2 Related Work

Personalized news recommendation plays a critical role in online news services to help users find their interested news information and is extensively studied over years (Konstan et al., 1997; Kompan and Bieliková, 2010; Wang and Blei, 2011; Bansal et al., 2015; Wu et al., 2021, 2019c; Lian et al., 2018; Khattar et al., 2018a). Existing personalized news recommendation methods usually first model user interests from historical news clicks, and then model the relevance between candidate news and user interests for personalized ranking. For example, Okura et al. (2017) proposed to learn user interest representation from clicked news sequence via a GRU network. They proposed to model the relevance between user interests and candidate news using the inner product of their representations. Wu et al. (2019e) proposed to learn user interest representations from user's clicked news with a multi-head self-attention network. They also used the inner product for matching user interest and candidate news from their representations. In these methods, user interests are modeled in a candidate-agnostic manner. However, users usually have multiple interests and it may be difficult to accurately match candidate news if candidate news information is not incorporated into user modeling. Different from these methods, *CAUM* incorporates a candidate-aware user modeling framework, which can use candidate news information to guide user modeling for better interest matching.

Only a few methods consider candidate news when modeling user interests (Wang et al., 2018; Zhu et al., 2019; Hu et al., 2020a). For example, Wang et al. (2018) proposed to use a candidate-aware attention network to model user interests by selecting clicked news based on their relevance to candidate news. Hu et al. (2020a) applied an LSTM network to process users' news click sequence and adopted a candidate-aware attention network to learn short-term user interest representations. They also build long-term user interest representations from a user-news interaction graph via a graph network. In these methods, candidate news information is only used in pooling the clicked news representation sequence into a unified user interest embedding via candidate-aware attention. However, candidate news information is not considered in modeling the contexts of news click behaviors, which may not be optimal for understanding user interests in candidate news. Different from these methods, our approach uses a candidate-aware self-attention network to model candidate-aware global user interests and a candidate-aware CNN network to model candidate-aware short-term user interests,

which can effectively evaluate user interests in candidate news for accurate news recommendation.

3 Methodology

In this section, we first present a formal definition of the problem studied in this paper, then introduce the details of our candidate-aware user modeling (*CAUM*) approach for news recommendation.

3.1 Problem Definition

Given a target user u and a set of candidate news $\{n_c^i | i = 1, 2, ..., M\}$, the task is to predict the matching score $\hat{y}_i$ measuring user interest on each candidate news n_c^i for personalized news ranking, where M represents the number of candidate news in the set. We assume that the user u has N historical clicked news, and we denote the i-th clicked news of this user as c_i. We assume that each news n contains three kinds of information, including news texts, news entities in news texts, and news topic category. We denote the i-th word in the news texts as w_i, the i-th entity as e_i, and the topic category as v. We denote the length of news texts as R and the number of entities as E.

3.2 Candidate-aware User Modeling

In general, users usually have multiple interests and a candidate news only matches a small part of user interests (Liu et al., 2020b). For example, Fig. 1 shows that the example user has multiple interests in different fields, including politics, music, sports, and travel. Besides, the first candidate news can match user interests in music, and it is irrelevant to other user interests. Thus, incorporating candidate news information into user modeling has the potential to match user interests with candidate news more accurately. Motivated by these observations, we propose a candidate-aware user modeling framework, which can exploit candidate news to guide user interests modeling. As shown in Fig. 2, it takes representations of user's clicked news $[\mathbf{c}_1, ..., \mathbf{c}_N]$, and representation of candidate news $\mathbf{n}_c$ as inputs, where $\mathbf{c}_i$ is the representation of the i-th click (News modeling method is introduced in section 3.3.). It contains three major modules, i.e., a candidate-aware self-attention network (*Candi-SelfAtt*), a candidate-aware CNN network (*Candi-CNN*), and a candidate-aware attention network (*Candi-Att*). We will introduce them in detail.

Candi-SelfAtt: Long-range contexts of news clicks are usually informative for inferring global user interests. For example, as shown in Fig. 1, the 1st click is a political news and the 5th click mentions Andrew Yang who is a politician. We can infer the user may be interested in political news mentioning Andrew Yang from the long-range relatedness between these two clicks. Besides, long-range behavior contexts usually have different importance to capture different global user interests. In Fig. 1, the relatedness between the 1st click and 5th click can help infer user interests in politics while the relatedness between the 5th click and 10th click can help infer user interests in music. Thus, modeling long-range behavior contexts with candidate news information may better model global user interests to match candidate news. Motivated by these observations, we propose a candidate-aware self-attention network (*Candi-SelfAtt*), which can use candidate news information to guide global behavior contexts modeling. The core of *Candi-SelfAtt* is to adjust attention weights of behavior contexts via candidate news to select important ones. First, we apply multiple self-attention heads (Vaswani et al., 2017) to model relatedness between the i-th click and other clicks:

$$\hat{r}_{i,j}^k = \mathbf{q}_i^T \mathbf{W}_r^k \mathbf{c}_j, \quad \mathbf{q}_i = \mathbf{Q}_u \mathbf{c}_i, \tag{1}$$

where $\hat{r}_{i,j}^k$ denotes the attention score generated by the k-th attention head, $\mathbf{q}_i$ is the query representation vector of the i-th clicked news, $\mathbf{Q}_u$ is the trainable projection matrix, and $\mathbf{W}_r^k$ is parameters of the k-th attention head. We further use candidate news information as guidance to select long-range contexts which are relevant to the candidate news:

$$r_{i,j}^k = \hat{r}_{i,j}^k + \mathbf{q}_c^T \mathbf{W}_r^k \mathbf{c}_j, \quad \mathbf{q}_c = \mathbf{Q}_c \mathbf{n}_c, \tag{2}$$

where $r_{i,j}^k$ is the candidate-aware attention score generated by the k-th self-attention head, $\mathbf{q}_c$ is the query vector of the candidate news, and $\mathbf{Q}_c$ is a trainable projection matrix. In this way, $\{r_{i,j}^k | j = 1, ..., N\}$ can encode candidate-aware long-range contexts between the i-th clicked news and other clicks. Then we learn contextual representation $\mathbf{l}_i^k$ generated by the k-th head for the i-th click:

$$\mathbf{l}_i^k = \mathbf{W}_o^k \sum_{j=1}^{N} \gamma_{i,j}^k \mathbf{c}_j, \quad \gamma_{i,j}^k = \frac{\exp(r_{i,j}^k)}{\sum_{p=1}^{N} \exp(r_{i,p}^k)}, \tag{3}$$

where $\gamma_{i,j}^k$ is the candidate-aware self-attention weight of $\mathbf{c}_j$ generated by the k-th self-attention head, and $\mathbf{W}_o^k$ is the projection matrix of the k-th

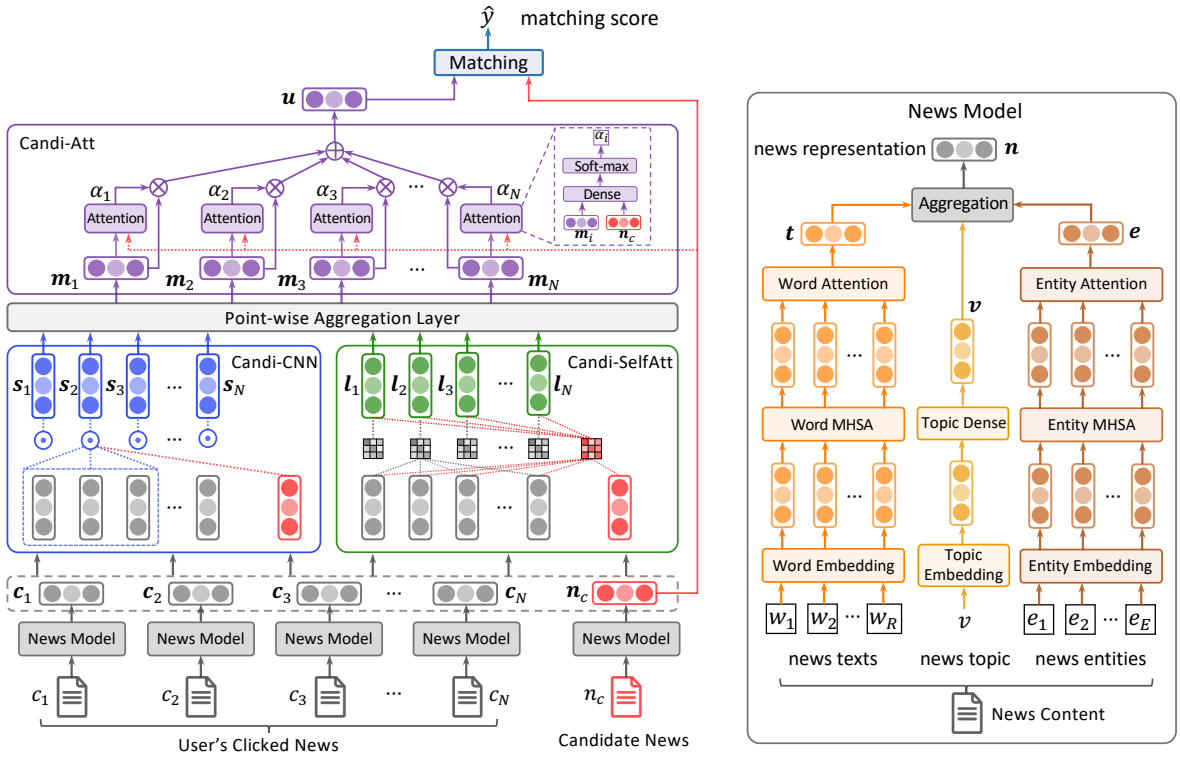

Figure 2: Framework of our *CAUM* method.

self-attention head. Finally, we learn the global contextual representation $\mathbf{l}_i$ for the i-th click by contacting its representations generated by all self-attention heads: $\mathbf{l}_i = [\mathbf{l}_i^1; \mathbf{l}_i^2; ...; \mathbf{l}_i^K]$, where K is the number of self-attention heads, $[\cdot;\cdot]$ represents the concatenation operation. Similarly, we can learn the global contextual representations for all user's clicked news $[\mathbf{l}_1, \mathbf{l}_2, ..., \mathbf{l}_N]$. These global contextual representations of news clicks can effectively encode candidate-aware global user interests.

Candi-CNN: Besides global user interests, short-term user interests are also important for matching candidate news (Hu et al., 2020a; An et al., 2019). Short-term user interests can usually be effectively modeled from local contexts between adjacent user behaviors (An et al., 2019). For example, as shown in Fig. 1, we can infer recent user interests on the travel in Wisconsin from local relatedness between 12th click and 13th click. Similarly, incorporating candidate news information into local behavior contexts modeling has the potential to better model short-term interest in candidate news. Thus, we propose a candidate-aware CNN network, which can capture local contexts between adjacent clicks with candidate news information. We apply multiple filters to capture the potential patterns between local contexts of adjacent clicks and candidate news:

$$\mathbf{s}_i = \mathbf{W}_c[\mathbf{c}_{i-h}; ...; \mathbf{c}_i; ...; \mathbf{c}_{i+h}; \mathbf{n}_c], \quad (4)$$

where $\mathbf{s}_i$ represents local contextual representation of the i-th click, $2h+1$ is the window size of the CNN network, and $\mathbf{W}_c$ represents parameters of filters in the *Candi-CNN* network. Similarly, we can learn local contextual representations $[\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_N]$ of all clicked news. These local contextual representations of clicked news encode candidate-aware short-term user interests. Then, we learn unified contextual representation $\mathbf{m}_i$ for the i-th click based on the aggregation of $\mathbf{l}_i$ and $\mathbf{s}_i$: $\mathbf{m}_i = \mathbf{P}_m[\mathbf{s}_i; \mathbf{l}_i]$, where $\mathbf{P}_m$ is the projection matrix.

Candi-Att: The importance of clicked news for modeling user interest in the candidate news are usually different. For example, Fig. 1 shows only 12th click and 13th click are informative for modeling user interests in travel, while other clicked news are not. Thus, we apply a candidate-aware attention network to model the importance of different clicked news based on their relevance with the candidate news n_c:

$$\alpha_i = \frac{\exp(\Phi(\mathbf{m}_i, \mathbf{n}_c))}{\sum_{j=1}^{N} \exp(\Phi(\mathbf{m}_j, \mathbf{n}_c))}, \quad (5)$$

where α_i is the attention weight of the i-th click, $\Phi(\cdot,\cdot)$ denotes a two-layer dense network for relevance measuring. We further learn the candidate-aware user interest representation $\mathbf{u}$ by aggregating contextual representations of clicks: $\mathbf{u}$ =

$\sum_{i=1}^{N} \alpha_i \mathbf{m}_i$. In this way, the candidate-aware user interest representation $\mathbf{u}$ can accurately model user interests for matching the candidate news n_c.

3.3 News Modeling

We design an effective news modeling method based on previous works since proposing a new news model is not the focus of our paper. As shown in Fig. 2, we model news content from news information of three views. First, we model news content from news texts. Motivated by Wu et al. (2019e), we apply a word embedding layer and a word multi-head self-attention (MHSA) network to model semantic information of news texts. Then we adopt a word attention network to learn text representation $\mathbf{t}$ for a news n. Second, we model news content from entities of news texts with the help of a knowledge graph. Following Liu et al. (2020a), we utilize an entity embedding layer to enhance knowledge information of the model and use an entity MHSA network to capture relatedness among entities. Besides, we apply an entity attention network to build entity representation $\mathbf{e}$. Third, we also exploit news topic v to better understand news content. Following Wu et al. (2019a), we derive embedding of news topic via a topic embedding layer and apply a topic dense network to learn topic representation $\mathbf{v}$. Finally we obtain news representation $\mathbf{n}$ by aggregating representations of different news information: $\mathbf{n} = \mathbf{W}_n[\mathbf{t}; \mathbf{e}; \mathbf{v}]$, where $\mathbf{W}_n$ is the projection matrix.

3.4 Interest Matching

Based on the news modeling and candidate-aware user modeling method, we can learn representation $\mathbf{n}_c$ of candidate news n_c, and the corresponding candidate-aware user interest representation $\mathbf{u}$. Following previous works (An et al., 2019; Wu et al., 2019e), we calculate the matching score $\hat{y}$ to measure user interest in candidate news via the inner product of their representations: $\hat{y} = \mathbf{n}_c \cdot \mathbf{u}$. The matching scores are further used to rank and recommend different candidate news.

3.5 Model Learning

Motivated by Wu et al. (2019d), we adopted BPR loss (Rendle et al., 2009) for model learning:

$$\mathcal{L} = -\frac{1}{H} \sum_{i=1}^{H} \log \phi(\hat{y}_i^p - \hat{y}_i^n), \qquad (6)$$

where $\mathcal{L}$ is the loss function, H is the size of training dataset, ϕ is the sigmoid function, $\hat{y}_i^p$ and $\hat{y}_i^n$ is the interest matching score of the i-th positive and negative sample. We randomly sample a negative sample (non-clicked news) for each positive sample (clicked news) from the same news impression.

	MIND	*NewsApp*
# News	161,013	1,126,508
# Users	1,000,000	50,605
# Topic Categories	18	28
Avg. # clicks of a user	24.2	19.4
Avg. # words in news title	11.78	11.90
Avg. # entities in news title	2.86	0.99

Table 1: Statistics of *MIND* and *NewsApp*.

4 Experiment

4.1 Dataset and Experimental Settings

We conduct extensive experiments on two real-world datasets to demonstrate the effectiveness of *CAUM*. The first one is *MIND*, a public dataset constructed by user logs on Microsoft News platform (Wu et al., 2020d). The second one is *NewsApp*, consisting of user logs collected from an anonymous news feeds app from January 23 to April 01, 2020 (13 weeks)[1]. It contains 100,000 and 10,000 impressions randomly selected from the first ten weeks to construct the training and validation set, and 100,000 impressions randomly selected from the last three weeks to construct the test set. We use clicks before time of news impression to construct users' click history. Table 1 shows more detailed information on these two datasets.

Next, we introduce all hyper-parameters of *CAUM* and experimental settings. For data processing, we use the first 30 words and 10 entities in news titles for modeling news content. Besides, we use the most recent 50 clicked news to model user interests. In *CAUM*, dimensions of both news and user interest representations are set to 400. For user modeling, *Candi-SelfAtt* contains 20 attention heads, and output vectors of each head are 20-dimensional. Besides, the query projection matrices, i.e., $\mathbf{Q}_u$ and $\mathbf{Q}_c$, generate 400-dimensional vectors. *Candi-CNN* contains 400 filters and window size is set to 3. *Candi-Att* is implemented by a two-layer dense network with 128-dimensional hidden vectors. For news modeling, 300-dimensional glove word embeddings, 100-dimensional TransE entity embeddings, and 100-dimensional random topic embeddings are used for initialization and fine-tuned in experiments. Word

[1]This dataset will be publicly released.

	MIND				*NewsApp*			
	AUC	MRR	nDCG@5	nDCG@10	AUC	MRR	nDCG@5	nDCG@10
GRU	65.69±0.15	31.47±0.06	33.96±0.07	39.70±0.07	63.23±0.37	27.83±0.26	31.84±0.31	37.41±0.34
NAML	66.49±0.19	32.38±0.13	35.17±0.15	40.84±0.14	64.52±0.35	29.02±0.20	33.35±0.30	38.90±0.33
NPA	66.56±0.18	32.42±0.10	35.20±0.11	40.87±0.13	64.39±0.14	28.93±0.10	33.31±0.11	38.83±0.11
NRMS	68.04±0.20	33.31±0.07	36.23±0.15	41.92±0.12	65.36±0.28	29.47±0.21	33.96±0.27	39.49±0.19
LSTUR	68.36±0.22	33.30±0.11	36.30±0.16	42.00±0.14	65.18±0.23	29.28±0.21	33.71±0.23	39.28±0.22
KRED	67.73±0.13	32.87±0.11	35.81±0.13	41.43±0.15	65.45±0.14	29.56±0.09	34.11±0.11	39.65±0.12
DKN	66.32±0.18	32.13±0.14	34.86±0.13	40.47±0.18	62.86±0.37	28.00±0.23	32.12±0.29	37.68±0.28
HiFi-Ark	67.93±0.25	32.87±0.07	35.77±0.08	41.47±0.10	64.91±0.15	29.10±0.12	33.52±0.18	38.98±0.14
FIM	67.84±0.12	33.26±0.06	36.18±0.10	41.86±0.11	65.39±0.10	29.63±0.11	34.14±0.12	39.60±0.10
GNewsRec	68.36±0.22	33.41±0.10	36.36±0.13	42.01±0.14	65.31±0.22	29.40±0.14	33.92±0.16	39.48±0.16
CAUM	**70.04**±0.08	**34.71**±0.08	**37.89**±0.07	**43.57**±0.07	**66.44**±0.07	**30.07**±0.10	**34.69**±0.12	**40.23**±0.10

Table 2: Performance of different methods on *MIND* and *NewsApp* datasets. T-test on these results verify that performance improvement of our *CAUM* method over other baseline methods is significant at level $p \leq 0.001$.

and entity transformer networks output 400- and 100-dimensional representations, respectively. We train *CAUM* 3 epochs via Adam (Kingma and Ba, 2015) with 5×10^{-5} learning rate. Besides, we apply dropout (Srivastava et al., 2014) to alleviate overfitting. All hyper-parameters of *CAUM* and other baseline methods are selected based on the validation dataset by manual tuning.[2] Following previous works (An et al., 2019; Wu et al., 2019a), we adopted AUC, MRR, nDCG@5, and nDCG@10 for evaluation.

4.2 Performance Comparison

We compare *CAUM* with several state-of-the-art baseline methods: (1) *GRU* (Okura et al., 2017): modeling user interests from user's clicked news via a GRU network (Cho et al., 2014). (2) *DKN* (Wang et al., 2018): proposing a candidate-aware attention network to learn user representations. (3) *NAML* (Wu et al., 2019a): building user representations via an attention network. (4) *NPA* (Wu et al., 2019b): proposing a personalized attention network to model user interests. (5) *HiFi-Ark* (Liu et al., 2019): learning user representations from multiple archives of user interests via a candidate-aware attention network. (6) *LSTUR* (An et al., 2019): modeling short-term user interests from user's recent clicked news via a GRU network and long-term user interests via user ID embeddings. (7) *NRMS* (Wu et al., 2019e): employing a multi-head self-attention network to learn user representations. (8) *KRED* (Liu et al., 2020a): modeling news content from news title and entities via a knowledge graph attention network. (9) *GNewsRec* (Hu et al., 2020a): modeling short-term user interests from clicked news sequence via a GRU network and a candidate-aware attention network, and long-term user interests from a user-news graph. (10) *FIM* (Wang et al., 2020): utilizing a CNN network (LeCun et al., 1998; Kim, 2014) to model user interests in candidate news from text similarities between clicked news and candidate news.

Each method is trained and evaluated 5 times. We list average performance and standard deviations in Table 2, from which we have several observations. First, we find that *CAUM* can significantly outperform other baseline methods which model user interests in a candidate-agnostic manner, such as *NRMS*, *LSTUR* and *KRED*. This is because users usually have multiple interests and a candidate news usually only matches a specific user interest. Modeling user interests in a candidate-agnostic manner makes these methods cannot effectively capture user interests that are relevant to a specific candidate news, and maybe sub-optimal for the interest matching. Different from these methods, *CAUM* exploits candidate news information to guide the modeling of user interests from clicked news and their contexts, which can better match user interests and candidate news.

Second, our *CAUM* method also outperforms baseline methods with candidate-aware attention network, such as *DKN*, *HiFi-Ark GNewsRec*. This is because different contexts of user's news clicks usually contain various clues to infer different user interests. Incorporating candidate news information into the behavior contexts modeling can help capture more relevant user interests for matching the candidate news. However, in these methods, user behavior contexts are ignored (e.g., *DKN*) or modeled in a candidate-agnostic way (e.g., *HiFi-Ark*), which are sub-optimal for modeling user interests in the candidate news. Different from these methods, in *CAUM* we propose a candidate-aware

[2]We uploaded our codes in the submission system and will publicly release them on GitHub.

	AUC	MRR	nDCG@5	nDCG@10
NAML	67.90±0.10	32.89±0.09	35.87±0.13	41.53±0.14
GRU	68.46±0.18	33.48±0.12	36.46±0.13	42.15±0.14
LSTUR	68.53±0.23	33.44±0.08	36.42±0.14	42.13±0.14
NRMS	68.46±0.20	33.42±0.07	36.45±0.10	42.13±0.13
DKN	68.07±0.16	33.47±0.09	36.50±0.12	42.16±0.10
HiFi-Ark	68.39±0.18	33.60±0.09	36.64±0.14	42.28±0.12
GNewsRec	68.37±0.19	33.31±0.06	36.27±0.08	41.96±0.11
CAUM	**70.04**±0.08	**34.71**±0.08	**37.89**±0.07	**43.57**±0.07

Table 3: Performance of user modeling methods.

self-attention network to use candidate news information as guidance to capture long-range contexts of user's historical clicks. Besides, we also propose a candidate-aware CNN network to capture local contexts of clicks with candidate news information.

4.3 Effectiveness of User Modeling

To compare *CAUM* with other user modeling methods more fairly, we evaluated their performance with the same news modeling method of *CAUM* introduced in section 3.3. Due to space limitation, we only show results on *MIND* in the following sections. As shown in Table 3, first we find that *CAUM* can significantly outperform baseline methods which model user interest in a candidate-agnostic manner. This further validates that modeling candidate-agnostic user interests is not optimal for interest matching. *CAUM* which uses candidate news information to guide user modeling can better match candidate news. Second, we find that *CAUM* significantly outperforms baseline methods with candidate-aware attention network. This further validates that incorporating candidate news information into behavior contexts modeling is beneficial for matching user interests and candidate news. *CAUM* can effectively use candidate news information as guidance to capture contexts of user's clicks via *Candi-SelfAtt* and *Candi-CNN*.

4.4 Ablation Study

We conduct an ablation study to verify the effectiveness of *Candi-SelfAtt* and *Candi-CNN* by adding them to the base model of *CAUM* (named *Base*). *Base* is a variation of *CAUM* that replaces *Candi-SelfAtt*, *Candi-CNN* and *Candi-Att* network with self-attention, CNN, and attention network individually. Results are shown in Fig. 3 and we have several findings. First, adding *Candi-SelfAtt* significantly improves the performance of *Base*. This is because different long-range behavior contexts usually contain various clues to infer different global user interests. However, the self-attention network models global user interests from behavior contexts

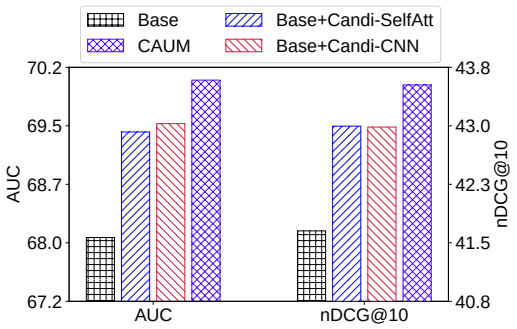

Figure 3: Ablation study of *CAUM*.

in a candidate-agnostic manner, which may be sub-optimal for further matching candidate news. Differently, *Candi-SelfAtt* can capture global user interests that are informative for matching the candidate news from long-range contexts of news clicks. Second, combing *Candi-CNN* with *Base* also has much better performance. Similarly, this is because modeling local behavior contexts with candidate news information is beneficial for short-term interest matching. CNN network is candidate-agnostics and maybe sub-optimal for modeling user interest in news. Differently, *Candi-CNN* can exploit candidate news information to capture local behavior contexts and model short-term user interest in candidate news. Third, *CAUM* outperforms both *Base+CandiCNN* and *Base+Candi-SelfAtt*. This is because user interest is usually composed of short- and long-term interest (An et al., 2019). *Candi-CNN* can only capture short-term user interest and *Candi-SelfAtt* mainly focus on capturing long-term user interest. Thus, combining them in *CAUM* can model user interest more accurately.

4.5 Analysis on Model Efficiency

We will present some efficiency analysis and comparisons on *CAUM* and other user modeling methods. First, in Table 4, we show time complexities of *CAUM* and candidate-agnostic methods for calculating matching scores of M candidate news for a user.[3] A notable result is that although *CAUM* needs to calculate different user representations for different candidate news, the time complexity of *CAUM* is not M times that of other methods. This is because, in *CAUM*, many operations only need to be performed once for different candidate news such as calculating self-attention scores $\hat{r}_{i,j}^{k}$ between different clicked news. Thus, by avoiding

[3]These methods can directly exploit news representations calculated in advance.

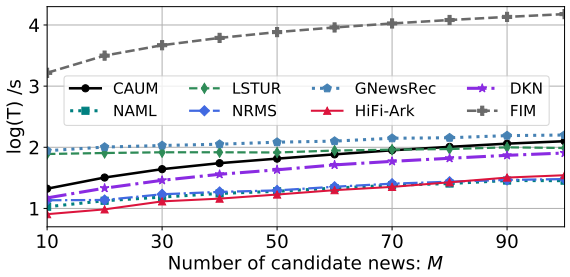

Figure 4: Efficiency comparisons of different methods.

NAML	$\mathcal{O}(Md + Nd^2)$	GRU	$\mathcal{O}(Md + Nd^2)$
LSTUR	$\mathcal{O}(Md + Nd^2)$	NRMS	$\mathcal{O}(3Nd^2 + N^2d + Md)$
CAUM	$\mathcal{O}((3N + M)d^2 + (N^2 + MN)d)$		

Table 4: Method time complexity (multiplication operation) of calculating matching scores of M candidate news. News and user representation are d-dimensional.

executing duplicated calculations, the efficiency of *CAUM* can be significantly improved. Besides, in general, the number of candidate news M is usually in a small scale (e.g., 100) in real-world recommender systems and it is comparable with the number of users' clicked news N used for interest modeling (e.g., 50)[4]. Thus, in practical settings, *CAUM* can achieve comparable time complexity with *NRMS*. In addition, although *GRU* and *LSTUR* have smaller time complexity than *NRMS* and *CAUM*, it is difficult to speed up these RNN based methods via parallel computations and they usually cost more time in real applications.

Second, as shown in Fig. 4, we compare running time T of different methods for calculating matching scores of M candidate news for 100,000 users. Different methods are executed in the same experimental environment (a Nvidia 1080 Ti GPU). We find that *CAUM* can achieve comparable speeds with many candidate-agnostic methods (e.g., *NAML* and *NRMS*) and outperform some candidate-agnostic methods (e.g., *LSTUR*). These results further verify that the efficiency of *CAUM* is satisfied like candidate-agnostic methods.

4.6 Case Study

As shown in Fig. 5, we conduct a case study to show the superiority of *CAUM* over candidate-agnostic methods. We compare the top 3 news recommended by *CAUM* and the most effective candidate-agnostic method in Table 3, i.e, *LSTUR* to a randomly sampled user in the same news impression. Fig. 5 shows that the user has multiple interests in different fields, such as sports, politics, and lifestyle. However, top news ranked by *LSTUR* are dominated by news on sports but the user did not click any of them. This is because *LSTUR* models user interests in a candidate-agnostic way and is hard to accurately match candidate news with a specific user interest. In addition, news recommended by *CAUM* can comprehensively cover user interests and the user clicked one of them. This is because *CAUM* can exploit candidate news information to guide the modeling of user interests, which is beneficial for accurate interest matching.

[4]In real applications candidate news are a small number of news recalled from a large-scale news database.

Reading history of a randomly sampled user		
1	sports	Tagovailoa, No. 1 Tide roll past No. 24 Texas A&M.
2	sports	With season halfway over, let's look at playoff picture.
3	lifestyle	Can you spot the camouflaged leopard in this picture?
4	politics	Trump was serious about chat to read Ukraine call.
5	sports	College football Week 11: Picks and preview.
Top 3 news recommended by LSTUR		
1	sports	Colin Kaepernick is about to get what he deserves.
2	sports	Bold predictions for Week 12 in college football.
3	sports	7 possible landing spots for Anthony Rendon.
Top 3 news recommended by CAUM		
1	sports	Bold predictions for Week 12 in college football.
2	politics	*4 takeaways from Marie Yovanovitch's testimony.*
3	lifestyle	Cows swept away by Hurricane Dorian found alive.

Figure 5: Case study. The clicked news in the randomly selected news impression is in purple and italic.

5 Conclusion

In this paper, we propose a candidate-aware user modeling framework for personalized news recommendation, which can incorporate candidate information into user modeling for more accurate interest matching. More specifically, we propose a candidate-aware self-attention network to exploit candidate news information as guidance to model global user interests in candidate news. Besides, we also propose a candidate-aware CNN network to incorporate candidate news information into local click behavior contexts modeling to match short-term user interests with the candidate news. In addition, we apply a candidate-aware attention network to build a unified user interest representation for matching candidate news by selecting important clicked news based on their relevance with candidate news. Extensive experiments on two real-world datasets validate the effectiveness of *CAUM* and demonstrate that our method can significantly outperform many baseline methods and improve the accuracy of user modeling.

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.

Trapit Bansal, Mrinal Das, and Chiranjib Bhattacharyya. 2015. Content driven user profiling for comment-worthy recommendations of news and blog articles. In *RecSys.*, pages 195–202.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *SSST*, pages 103–111.

Sandra Garcia Esparza, Michael P O'Mahony, and Barry Smyth. 2010. On the real-time web as a source of recommendation knowledge. In *RecSys.*, pages 305–308.

Suyu Ge, Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2020. Graph enhanced representation learning for news recommendation. In *WWW*, pages 2863–2869.

Linmei Hu, Chen Li, Chuan Shi, Cheng Yang, and Chao Shao. 2020a. Graph neural news recommendation with long-term and short-term interest modeling. *IP&M*, page 102142.

Linmei Hu, Siyong Xu, Chen Li, Cheng Yang, Chuan Shi, Nan Duan, Xing Xie, and Ming Zhou. 2020b. Graph neural news recommendation with unsupervised preference disentanglement. In *ACL*, pages 4255–4264.

Dhruv Khattar, Vaibhav Kumar, Vasudeva Varma, and Manish Gupta. 2018a. Hram: A hybrid recurrent attention machine for news recommendation. In *CIKM*, pages 1619–1622.

Dhruv Khattar, Vaibhav Kumar, Vasudeva Varma, and Manish Gupta. 2018b. Weave&rec: A word embedding based 3-d convolutional network for news recommendation. In *CIKM*, pages 1855–1858.

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

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Michal Kompan and Mária Bieliková. 2010. Content-based news recommendation. In *EC-Web*, pages 61–72.

Joseph A Konstan, Bradley N Miller, David Maltz, Jonathan L Herlocker, Lee R Gordon, and John Riedl. 1997. Grouplens: applying collaborative filtering to usenet news. *Communications of the ACM*, pages 77–87.

Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, pages 2278–2324.

Dongho Lee, Byungkook Oh, Seungmin Seo, and Kyong-Ho Lee. 2020. News recommendation with topic-enriched knowledge graphs. In *CIKM*, pages 695–704.

Jianxun Lian, Fuzheng Zhang, Xing Xie, and Guangzhong Sun. 2018. Towards better representation learning for personalized news recommendation: a multi-channel deep fusion approach. In *IJCAI*, pages 3805–3811.

Chen Lin, Runquan Xie, Xinjun Guan, Lei Li, and Tao Li. 2014. Personalized news recommendation via implicit social experts. *JIS*, pages 1–18.

Danyang Liu, Jianxun Lian, Shiyin Wang, Ying Qiao, Jiun-Hung Chen, Guangzhong Sun, and Xing Xie. 2020a. Kred: Knowledge-aware document representation for news recommendations. In *RecSys.*, pages 200–209.

Zheng Liu, Jianxun Lian, Junhan Yang, Defu Lian, and Xing Xie. 2020b. Octopus: Comprehensive and elastic user representation for the generation of recommendation candidates. In *SIGIR*, pages 289–298.

Zheng Liu, Yu Xing, Fangzhao Wu, Mingxiao An, and Xing Xie. 2019. Hi-fi ark: deep user representation via high-fidelity archive network. In *IJCAI*, pages 3059–3065.

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

Tao Qi, Fangzhao Wu, Chuhan Wu, Yongfeng Huang, and Xing Xie. 2020. Privacy-preserving news recommendation model learning. In *EMNLP: Findings*, pages 1423–1432.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. Bpr: Bayesian personalized ranking from implicit feedback. In *UAI*, pages 452–461.

TYSS Santosh, Avirup Saha, and Niloy Ganguly. 2020. Mvl: Multi-view learning for news recommendation. In *SIGIR*, pages 1873–1876.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, pages 1929–1958.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 6000–6010.

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

Heyuan Wang, Fangzhao Wu, Zheng Liu, and Xing Xie. 2020. Fine-grained interest matching for neural news recommendation. In *ACL*, pages 836–845.

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

Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019a. Neural news recommendation with attentive multi-view learning. *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, Mingxiao An, Yongfeng Huang, and Xing Xie. 2019c. Neural news recommendation with topic-aware news representation. In *ACL*, pages 1154–1159.

Chuhan Wu, Fangzhao Wu, Mingxiao An, Tao Qi, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019d. Neural news recommendation with heterogeneous user behavior. In *EMNLP*, pages 4876–4885.

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

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2020a. Sentirec: Sentiment diversity-aware neural news recommendation. In *AACL*, pages 44–53.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2020b. User modeling with click preference and reading satisfaction for news recommendation. In *IJCAI*, pages 3023–3029.

Chuhan Wu, Fangzhao Wu, Tao Qi, Jianxun Lian, Yongfeng Huang, and Xing Xie. 2020c. Ptum: Pre-training user model from unlabeled user behaviors via self-supervision. In *EMNLP: Findings*, pages 1939–1944.

Chuhan Wu, Fangzhao Wu, Xiting Wang, Yongfeng Huang, and Xing Xie. 2021. Fairrec:fairness-aware news recommendation with decomposed adversarial learning. In *AAAI*.

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. 2020d. Mind: A large-scale dataset for news recommendation. In *ACL*, pages 3597–3606.

Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie, and Zhenhui Li. 2018. Drn: A deep reinforcement learning framework for news recommendation. In *WWW*, pages 167–176.

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