Neural News Recommendation with Heterogeneous User Behavior

Chuhan Wu[1], Fangzhao Wu[2], Mingxiao An[3], Tao Qi[1], Jianqiang Huang[4], Yongfeng Huang[1], and Xing Xie[2]
[1]Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[2]Microsoft Research Asia, Beijing 100080, China
[3]University of Science and Technology of China, Hefei 230026, China
[4]Peking University, Beijing 100871, China
{wu-ch19, qit16, yfhuang}@mails.tsinghua.edu.cn,
{fangzwu, xing.xie}@microsoft.com,
anmx@mail.ustc.edu.cn, 1701210864@pku.edu.cn

Abstract

News recommendation is important for online news platforms to help users find interested news and alleviate information overload. Existing news recommendation methods usually rely on the news click history to model user interest. However, these methods may suffer from the data sparsity problem, since the news click behaviors of many users in online news platforms are usually very limited. Fortunately, some other kinds of user behaviors such as webpage browsing and search queries can also provide useful clues of users' news reading interest. In this paper, we propose a neural news recommendation approach which can exploit heterogeneous user behaviors. Our approach contains two major modules, i.e., news representation and user representation. In the news representation module, we learn representations of news from their titles via CNN networks, and apply attention networks to select important words. In the user representation module, we propose an attentive multi-view learning framework to learn unified representations of users from their heterogeneous behaviors such as search queries, clicked news and browsed webpages. In addition, we use word- and record-level attentions to select informative words and behavior records. Experiments on a real-world dataset validate the effectiveness of our approach.

1 Introduction

Online news platforms such as Google News[1] and MSN[2] News have gained huge popularity for online digital news reading (Das et al., 2007). Tens of thousands of news articles are streamed from various sources every day, making it very difficult for users to read all news to find their interested content (Phelan et al., 2011). Thus, personalized news recommendation is critical for online news platforms to target user interests and alleviate information overload (IJntema et al., 2010).

Figure 1: An illustrative example of news recommendation based on heterogeneous user behaviors.

News and user representation learning is critical for news recommendation. Many deep learning based methods have been proposed for this task (Okura et al., 2017; Wang et al., 2018). For example, Okura et al. (2017) proposed to learn news representations from news bodies via auto-encoders, and learn user representations from their clicked news via a gated recurrent unit (GRU) network. Wang et al. (2018) proposed to learn news representations from news titles via a knowledge-aware convolutional neural network (CNN), and learn user representations from their clicked news using the similarities between candidate news and clicked news. These methods rely on the news click histories to model users' news reading interest. However, these methods usually suffer from the data sparsity problem, since the news click behaviors of many users on news platforms are very limited, making it difficult for these methods to learn accurate representations of these users.

Luckily, there are several other kinds of user

[1]https://news.google.com/
[2]https://www.msn.com/en-us/news

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 4874–4883,
Hong Kong, China, November 3–7, 2019.

behaviors such as search queries and webpage browsing which are useful for news recommendation. Many online users frequently use search engines such as Google and Bing to search for desired information (Wu et al.). In addition, they may browse related webpages to get more detailed information. Thus, the search and browsing data accumulated by commercial search engines can cover a large number of online users. In addition, the search and browsing behaviors of users can provide rich information for inferring user interests. For example, in Fig. 1, since the user posts the search query "toyota crown", and browses the webpage of "Pros and Cons of Toyota" for detailed information, she may have high interest in Toyota cars and is likely to read news articles related to Toyota cars. However, this user may click only a few news articles, and her interest in toyota cars cannot be mined from the news click history. Thus, incorporating the heterogeneous behaviors of online users such as search queries and webpage browsing has the potential to improve the performance of news recommendation.

Our work is motivated by following observations. First, the characteristics of search queries, webpage titles and news have huge differences. For example, in Fig. 1, search queries are usually phrase pieces with a few words, while news and webpage titles are usually complete sentences. Thus, they should be handled differently. Second, different words usually have different importance in representing behavior records. For example, in Fig. 1 the word "NFL" is more informative than "Today". Third, different behavior records may also have different importance for representing users. For instance, the search query "safe cars" in Fig. 1 is more informative than the query "google map" in modeling user interest. Fourth, different kinds of user behaviors usually have different informativeness for user representation learning. For example, in Fig. 1 since the user clicks very few news articles, her news click behaviors are not very informative for modeling this user.

In this paper, we propose a neural news recommendation approach with heterogeneous user behavior (NRHUB). The core of our approach is a news representation module and a user representation module. In the news representation module, we learn news representations from news titles via a CNN network with word-level attentions to select important words. In the user representation module, we propose an attentive multi-view learning framework to incorporate the different kinds of user behavior data into our model to learn unified user representations. In each view, we propose to use hierarchical user encoders to first learn record representations from words, and then learn user representations from different kinds of behavior records. Since different views may have different informativeness for modeling users, we propose to use a view-level attention network to select important views for learning informative user representations. In addition, since different words and records may also have different informativeness, we apply word-level and record-level attention networks to select important words and records. Extensive experiments are conducted on a real-world news recommendation dataset. The results show that our approach can effectively improve the performance of news recommendation, especially in the cold-start scenario.

2 Related Work

News recommendation is an important task in both natural language processing and data mining fields, and has wide applications (Okura et al., 2017; Zheng et al., 2018). Learning accurate news and user representations is critical for news recommendation. Many of existing methods for news recommendation rely on manual feature engineering to build news and user representations (Liu et al., 2010; Son et al., 2013; Bansal et al., 2015). For example, Liu et al. (2010) proposed to use topic categories and interest features generated by a Bayesian model to build news and user representations. Son et al. (2013) proposed an Explicit Localized Semantic Analysis (ELSA) model for location-based news recommendation. They proposed to extract topic and location features from Wikipedia pages to build news representations. Lian et al. (2018) proposed a deep fusion model (DMF) to learn news representations from various handcrafted features such as title length, entities and news topics, and learn user representations from features extracted from the news reading, web browsing, and searching behaviors of users. However, these methods usually rely on manual feature engineering, which necessities massive domain knowledge and time to design. In addition, these methods cannot exploit contextual information in the behavior records of users, which may be insufficient to learn accurate user representations.

In recent years, several deep learning based methods are proposed for news recommendation (Okura et al., 2017; Wang et al., 2018; Khattar et al., 2018; Kumar et al., 2017; Zheng et al., 2018; Wu et al., 2019b; An et al., 2019; Wu et al., 2019c,a; Zhu et al., 2019). For example, Okura et al. (2017) proposed to learn news representations from news bodies via denoising autoencoders, and learn user representations from the news browsing sequence via a GRU network. Wang et al. (2018) proposed to learn news representations from news titles via a knowledge-aware CNN network, which can incorporate useful entity information from knowledge graphs. Wu et al. (2019b) proposed to learn news representations from news titles, and apply personalized attention mechanism at both word- and news-level to generate representations of news and users according to user preferences. However, these methods only learn user representations from a single kind of user behavior, i.e., news click, which may be insufficient. Different from these methods, our approach can learn unified user representations by incorporating heterogeneous user behaviors via an attentive multi-view learning framework. Extensive experiments on real-world dataset validate our approach can learn better news and user representations, and achieve better performance on news recommendation than existing methods.

3 Our Approach

In this section, we introduce our neural news recommendation approach with heterogeneous user behavior (NRHUB) in detail. The architecture of our *NRHUB* approach is shown in Fig. 2. Our approach has three major modules, i.e., *news representation*, *user representation* and *click predictor*. Next, we introduce the details of each module.

3.1 News Representation Learning

The *news representation* module is used to learn representations of news articles from their titles. It contains three layers.

The first one is a word embedding layer, which is used to convert a news title from a sequence of words into a sequence of low-dimensional semantic vectors. Denote a news title with M words as $[w_1, w_2, ..., w_M]$. It is converted into a vector sequence $[\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_M]$ via a pre-trained embedding matrix.

The second one is a CNN layer (Kim, 2014). Usually, local contexts are very important in representing the news titles. For example, in the news title "Toyota Crown Classics for Sale", local contexts of "Crown" such as "Toyota" are useful for inferring the topic of this news, i.e., cars. Thus, we apply a CNN network to learn contextual representations of words within news titles by capturing their contexts. The CNN layer computes the contextual representations of each word, and outputs a vector sequence $[\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_M]$.

The third one is a word-level attention network. Different words in the same news title may have different importance in representing this news. For example, in the second news in Fig. 1, the word "NFL" is more informative than "Today" for representing this news. Thus, we propose to use attention mechanism to select important words in news titles for learning informative news representations. Denote the attention weight of the i-th word in a news title as α_i^w:

$$
\begin{aligned}
a_i^w &= \mathbf{v}_w^n \times \tanh(\mathbf{V}_w \times \mathbf{c}_i + \mathbf{v}_w), \\
\alpha_i^w &= \frac{\exp(a_i^w)}{\sum_{j=1}^{M} \exp(a_j^w)},
\end{aligned} \tag{1}
$$

where $\mathbf{V}_w$ and $\mathbf{v}_w$ are projection parameters, $\mathbf{v}_w^n$ is the query vector in this attention network. The final contextual representation of a news title is the summation of its word representations weighted by their attention weights as follows:

$$
\mathbf{r}^n = \sum_{i=1}^{M} \alpha_i^w \mathbf{c}_i. \tag{2}
$$

3.2 User Representation Learning

The *user representation* module is used to learn the representations of users from different kinds of user behaviors, e.g., web searching, news click and webpage browsing. Since these user behaviors usually have different characteristics, simply aggregate them together into a long document may not be optimal for news representation. Thus, we propose an attentive multi-view learning framework to learn unified user representations by incorporating these three kinds of user behavior information as different views of news.

The first view is *news encoder*, which is used to learn user representations from his/her clicked news. There are two major submodules in this view. The first one is news representation, as described in section 3.1. Denote the sequence of news clicked by a user as $[D_1, D_2, ..., D_N]$, where

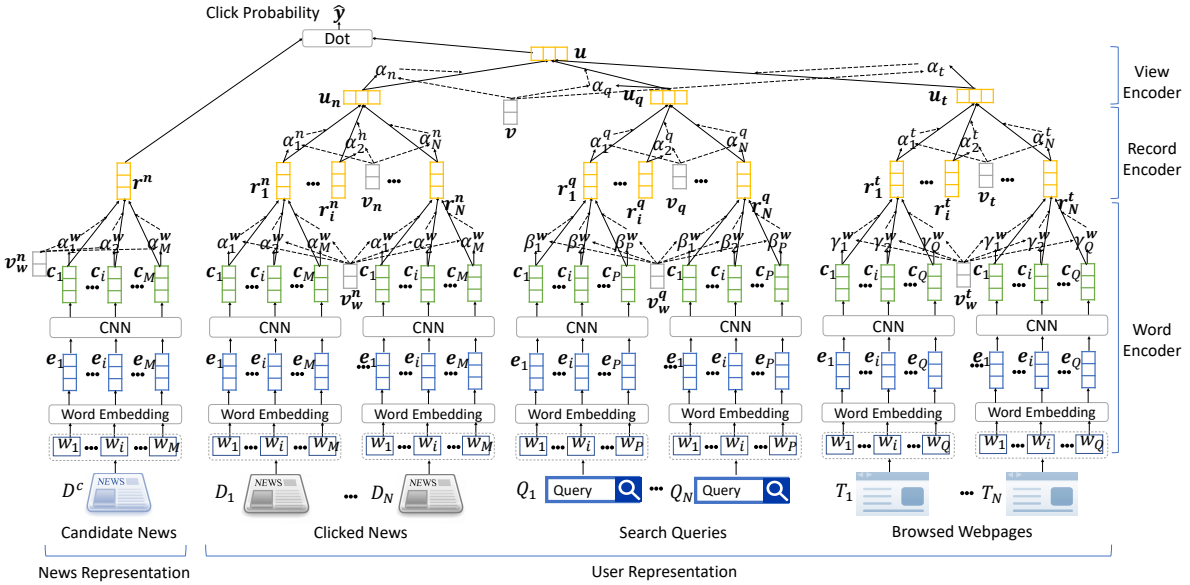

Figure 2: The framework of our *NRHUB* approach.

N is the number of clicked news articles. It is transformed into a sequence of hidden news representations, denoted as $[\mathbf{r}_1^n, \mathbf{r}_2^n, ..., \mathbf{r}_N^n]$.

The second one is a news-level attention network. Different news clicked by the same user may have different informativeness for representing this user. For example, in Fig. 1, the second news is more informative than the first news in modeling user preferences, since the first one is usually not audience sensitive. Thus, we use a news-level attention network to select important news for learning informative user representations. Denote the attention weight of the i-th news clicked by a user as α_i^n, which is computed as:

$$a_i^n = \mathbf{v}_n \times \tanh(\mathbf{V}_n \times \mathbf{r}_i + \mathbf{b}_n),$$
$$\alpha_i^n = \frac{\exp(a_i^n)}{\sum_{j=1}^{N} \exp(a_j^n)}, \tag{3}$$

where $\mathbf{V}_n$ and $\mathbf{b}_n$ are parameters, $\mathbf{v}_n$ is the attention query vector. The news based representation of a user u is the summation of the representations of clicked news articles weighted by their attention weights as follows:

$$\mathbf{u}_n = \sum_{i=1}^{N} \alpha_i^n \mathbf{r}_i^n. \tag{4}$$

The second view is *query encoder*, which is used to learn representations of users from their search queries. There are two submodules in this view. The first module is query representation, which is used to learn query representations from words. It also has three layers.

The first one is a word embedding layer, which is shared with the news representation view. Denote a search query q with P words as $q = [w_1, w_2, ..., w_P]$. It is converted into a sequence of hidden word representations, denoted as $[\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_P]$.

The second one is a CNN layer. Local contexts are very important in learning query representations. For example, in the query "xbox one player", the local contexts of "one", i.e., "xbox" and "player" are important in representing this query. Thus, we apply CNN to capture the local contexts within queries, and compute the contextual representations of each word. The outputs of this CNN layer is denoted as $[\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_P]$.

The third one is a word-level attention network. Different words in the same search query usually have different informativeness for representation learning. For example, in the query "hot movies this week" "movies" is more informative than "week" in modeling this query. Thus, we propose to use a word-level attention network to recognize important words in search queries for learning informative query representations. The attention weight β_i^w of the i-th word in a search query is computed as:

$$b_i^w = \mathbf{v}_w^q \times \tanh(\mathbf{V}_q \times \mathbf{c}_i + \mathbf{v}_q),$$
$$\beta_i^w = \frac{\exp(b_i^w)}{\sum_{j=1}^{P} \exp(b_j^w)}, \tag{5}$$

where $\mathbf{V}_q$, $\mathbf{v}_q$ and $\mathbf{v}_w^q$ are parameters in the attention network. The contextual representation of this query is the summation of its word representations weighted by their attention weights, as follows:

$$\mathbf{r}^q = \sum_{i=1}^{P} \beta_i^w \mathbf{c}_i. \tag{6}$$

The second module is a query-level attention network, which is used to learn user representations from query representations. Different queries posted by the same user usually have different importance in characterizing this user. For example, the query "toyota cars" is more informative than the query "google mail", since the later one does not contain information of user preference. To select important search queries for learning informative user representations, we propose to incorporate a query-level attention network. The attention weight α_i^q of the query Q_i is computed as:

$$\begin{aligned} a_i^q &= \mathbf{v}_q \times \tanh(\mathbf{U}_q \times \mathbf{r}_i + \mathbf{u}_q), \\ \alpha_i^q &= \frac{\exp(a_i^q)}{\sum_{j=1}^{N} \exp(a_j^q)}, \end{aligned} \tag{7}$$

where $\mathbf{U}_q$, $\mathbf{u}_q$ and $\mathbf{v}_q$ are attention parameters. The user representation learned from queries is the summation of the query representations weighted by their attention weights as:

$$\mathbf{u}_q = \sum_{i=1}^{N} \alpha_i^q \mathbf{r}_i^q. \tag{8}$$

The third view is *webpage encoder*. It is used to learn representations of users from the titles of their browsed webpages. There are two submodules in this view. The first one is *webpage representation*, which is used to learn the representations of webpage titles from their words, and the second one is a webpage-level attention network which learns user representation from webpages. In the *webpage representation* module, we first use a CNN network to learn contextual representations of the words in webpage titles, and then we use a word-level attention network to highlight important words. The hidden representation of each webpage is the summation of the contextual representations of words weighted by their attention weights, which is similar to the search query and news views described previously. The webpage-level attention network is used to select informative webpages for user representation learning, and the final webpage based user representation $\mathbf{u}_t$ is the summation of the webpage representations weighted by their attention weights.

The last component in the *user representation* module is a view encoder. Usually, different views may have different informativeness for modeling users. For example, a user may never or rarely cl news, and his/her news click data may be uninformative for representing this user. Thus, modeling the informativeness of different views may benefit user representation learning. Motivated by these observations, we propose a view-level attention network to select important views for learning informative user representations. Denote the attention weights of the news, query and webpage views as α_n, α_q and α_t, respectively. For instance, the attention weight of the news title view is computed as follows:

$$\begin{aligned} a_n &= \mathbf{v} \times \tanh(\mathbf{U}_v \times \mathbf{u}_n + \mathbf{u}_v), \\ \alpha_n &= \frac{\exp(a_n)}{\exp(a_n) + \exp(a_q) + \exp(a_t)}, \end{aligned} \tag{9}$$

where $\mathbf{v}$, $\mathbf{U}_v$ and $\mathbf{u}_v$ are parameters. The attention weights of the search query and webpage title views are computed similarly. The unified user representation is the summation of the user representations from different views weighted by their attention weights, which is formulated as

$$\mathbf{u} = \alpha_n \mathbf{u}_n + \alpha_q \mathbf{u}_q + \alpha_t \mathbf{u}_t. \tag{10}$$

3.3 Click Predictor

The *click predictor* module is used to predict the probability of a user clicking a candidate news article from their representations. Denote the representation of the candidate news D^c as $\mathbf{r}^n$. Following (Okura et al., 2017), the click probability score $\hat{y}$ is computed by the inner product of the user and candidate news representations, i.e., $\hat{y} = \mathbf{u}^T \mathbf{r}^n$.

3.4 Model Training

Motivated by (Huang et al., 2013), we propose to use the negative sampling technique for model training. For each news clicked by a user which is regarded as a positive sample, we randomly sample K news in the same impression but not clicked by this user as the negative samples. Denote the click probability score of the positive news as $\hat{y}^+$, and scores of the K negative news as $[\hat{y}_1^-, \hat{y}_2^-, ..., \hat{y}_K^-]$. We normalize these scores using the softmax function to compute the posterior

# users	10,000	**avg. # words per new title**	11.29
# news	42,255	**avg. # displays per news**	247.89
# queries	923,041	**# user w/ queries**	5,282
# webpage titles	813,852	**# user w/ webpages**	6,134
# impressions	360,428	**avg. # words per query**	3.25
# samples	10,474,493	**avg. # words per webpage title**	9.01
# positive samples	503,698	**# negative samples**	9,970,795

Table 1: Statistics of our dataset.

click probability of a positive sample as follows:

$$p_i = \frac{\exp(\hat{y}_i^+)}{\exp(\hat{y}_i^+) + \sum_{j=1}^{K} \exp(\hat{y}_{i,j}^-)}. \quad (11)$$

The loss function for model training is the negative log-likelihood of all positive samples:

$$\mathcal{L} = -\frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \log(p_i), \quad (12)$$

where $\mathcal{S}$ is the set of positive training samples.

4 Experiments

4.1 Datasets and Experimental Settings

We conducted experiments on a real-world news recommendation dataset[3] collected from MSN News[4] logs during Dec. 13, 2018 and Jan. 12, 2019. In addition, we crawled the search queries and titles of browsed webpages from the logs of the Bing search engine. The detailed statistics of this dataset are summarized in Table 1. The news data in the last week is used for test, and the rest is used for model training. In addition, we randomly sampled 10% of the training data for validation.

In our experiments, the word embeddings were 300-dimensional and we used the pre-trained Glove embedding for initialization (Pennington et al., 2014). All CNN networks had 400 filters, and their window size was set to 3. The dimension of attention query vectors was 200. Following (Wu et al., 2019b), the negative sampling ratio K was set to 4. Adam (Kingma and Ba, 2014) was used as the optimization algorithm. We applied 30% dropout to the word embedding and CNN networks to mitigate overfitting. The training batch size was 64. These hyperparameters were all tuned on the validation set. We independently repeated each experiment 10 times and reported the average results in AUC, MRR, nDCG@5 and nDCG@10 of all impressions.

4.2 Performance Evaluation

We evaluate the performance of our *NRHUB* approach by comparing it with many baseline methods, including: (1) *LibFM* (Rendle, 2012), a matrix factorization method for recommendation; (2) *DSSM* (Huang et al., 2013), deep structured semantic model; (3) *Wide&Deep* (Cheng et al., 2016), a popular neural recommendation method with a wide linear model and a deep neural network; (4) *DeepFM* (Guo et al., 2017), a famous neural recommendation method with a factorization machine and a deep neural network; (5) *DFM* (Lian et al., 2018), a deep fusion model for news recommendation; (6) *GRU* (Okura et al., 2017), a neural news recommendation method based on autoencoder and GRU to learn news and user representations; (7) *DKN* (Wang et al., 2018), a deep knowledge-aware news recommendation method; (8) *Conv3D* (Kumar et al., 2017; Khattar et al., 2018), a nerual news recommendation method based on a 3-D convolutional neural network to learn user representations from the sequence of clicked news[5]; (9) *NRHUB*, our neural new recommendation approach with heterogeneous user behavior; (10) *NRHUB-basic*, a variant of our approach with only news click behavior; (11) *NRHUB-concat*, a variant of our approach which aggregates different kinds of user behaviors as a single view. Following (Wang et al., 2018), in methods (1, 3-5), we use one-hot encoded user ID, news ID and the TF-IDF features of news titles as the input features. The results are summarized in Table 2.

From Table 2, we have several observations. First, deep learning based news recommendation methods (e.g., *DSSM* and *NRHUB*) can outperform *LibFM*. This is because neural networks can

[3]Some publicly available resources can be found at https://github.com/wuch15/NRHUB.

[4]https://www.msn.com/en-us/news

[5]We only use news titles for fair comparison.

Methods	AUC	MRR	nDCG@5	nDCG@10
LibFM	0.5661	0.2414	0.2689	0.3552
DSSM	0.5949	0.2675	0.2881	0.3800
Wide & Deep	0.5812	0.2546	0.2765	0.3674
DeepFM	0.5830	0.2570	0.2802	0.3707
DFM	0.5861	0.2609	0.2844	0.3742
GRU	0.6114	0.2823	0.3006	0.3931
DKN	0.6070	0.2801	0.3012	0.3933
Conv3D	0.6051	0.2765	0.2987	0.3904
NRHUB-basic	0.6232	0.2927	0.3158	0.3986
NRHUB-concat	0.6265	0.2945	0.3177	0.4010
NRHUB*	**0.6317**	**0.3020**	**0.3260**	**0.4076**

Table 2: The results of different methods. *The improvement is significant at $p < 0.01$.

learn better news and user representations than traditional matrix factorization methods, which is beneficial for more accurate news recommendation. Second, among deep learning based methods, both *NRHUB-basic* and *NRHUB* can outperform all baseline methods. This is because our approaches incorporate both word- and news-level attention networks to simultaneously select important words and news for learning more informative news and user representations. Thus, our approaches can outperform baseline methods. Third, *NRHUB* and *NRHUB-concat* methods outperform *NRHUB-basic*. This is because search queries and browsed webpage can provide rich information of user interests, which is useful for learning user representation for news recommendation. Fourth, *NRHUB* consistently outperforms *NRHUB-concat*. This is because the characteristics of news titles, search queries and webpage titles are quite different, and they should be handled differently. Thus, simply merging them together is not the optimal way to exploit them. These results validate the effectiveness of our *NRHUB* approach in incorporating heterogeneous user behaviors for news recommendation.

4.3 Effectiveness of Attentive Multi-view Learning

In this section, we explore the effectiveness of the attentive multi-view learning framework in our approach. First, we want to verify the effectiveness of incorporating heterogeneous user behaviors as different views. We compare the performance of our *NRHUB* approach with its variants with different combinations of views, and the results are shown in Fig. 3. From Fig. 3, we have several observations. First, both search queries and browsed webpage titles are useful for improving the performance of news recommendation. This may be because users usually use post search queries to search engines to seek for interested content, and may browse several related webpages for detailed information. Thus, search queries and browsed webpage titles contain rich information of user preferences, which is useful for learning accurate user representations. Second, webpage titles are more important than search queries. This may be because webpage titles have larger coverage in our dataset. In addition, search queries are usually very short with a few words, while webpage titles are often like complete sentences with much more words. Thus, webpage titles are more informative for learning user representations. Third, incorporating both kinds of user behaviors can further improve the performance of our approach. These results validate the effectiveness of incorporating heterogeneous user behaviors via multi-view learning.

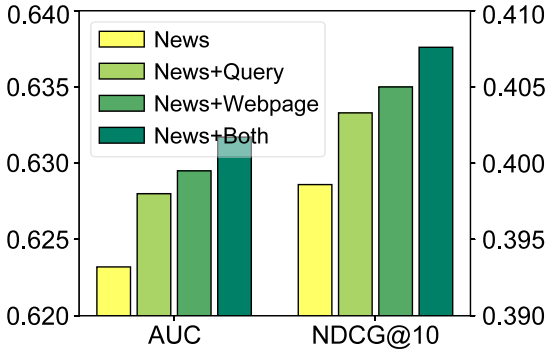

Figure 3: Effect of heterogeneous user behaviors.

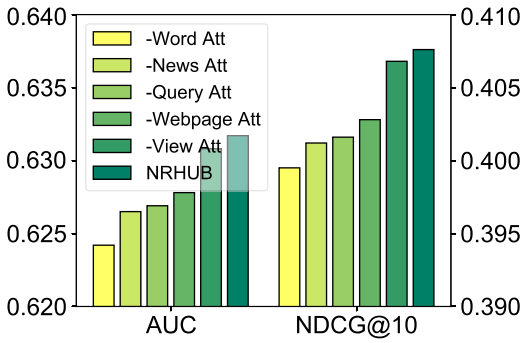

Figure 4: Influence of different attention networks.

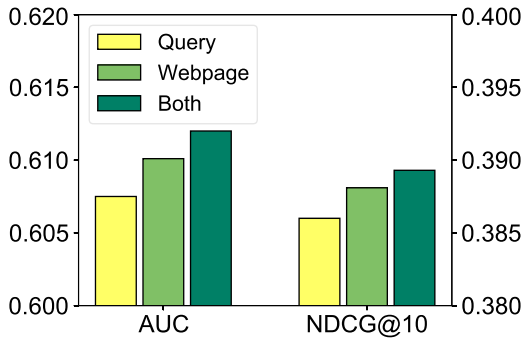

Figure 5: Cold-start performance of our approach.

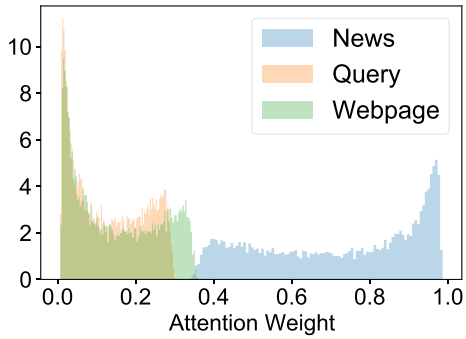

Figure 6: Distributions of view-level attention weights. The left, middle and right curves respectively denote search query, browsed webpage and clicked news.

Next, we explore the effectiveness of different attention networks in our approach. We conduct experiments using the leave-one-out scheme to evaluate the contribution of each attention network. The results are shown in Fig. 4. From Fig. 4, we find the word-level attention network can effectively improve the performance of our approach. This is because different words in the same news, query or webpage title may have different importance for learning user representations. Thus, selecting important words can help learn more informative news, query and webpage representations. In addition, the news-, query- and webpage-level attention networks are also important to our approach. This is because different news articles, search queries or webpages usually have different importance in representing users, and selecting important news, queries and webpages is beneficial for learning accurate user representations. Besides, the view-level attention network is also useful. This is because different views also have different informativeness in modeling user preferences. Thus, evaluating the informativeness of different views via a view-level attention network can learn better user representations. Moreover, combining all these kinds of attention networks can further improve the performance of our approach. These results validate the effectiveness of the attentive multi-view learning framework in our approach.

4.4 Cold-start Performance

In this section, we explore the cold-start performance of our approach. In practical use, massive users may rarely or have never clicked news articles on the target news platform, making it difficult to effectively recommend news to these users. Luckily, the search and browsing behaviors can cover many of these users. These user behaviors usually contain rich information of user preferences, and have the potential to address the cold-start problem. To validate the effectiveness of our approach, we compare the performance of several variants of our approach without the news encoder view to simulate the cold-start situation. The results are shown in Fig. 5. According to the results, we find our approach can still achieve satisfactory performance in the cold-start situation. This may be because search queries and webpage titles are important clues of user interests, which are very useful for news recommendation. Thus, our approach can recommend news effectively to the users with very limited or even no news browsing histories, which can effectively improve their experiences. In addition, combining the information of queries and webpage titles can further improve the performance of our approach. These results validate the effectiveness of our approach.

4.5 Case Study

In this section, we conduct several case studies to visually explore the effectiveness of different at-

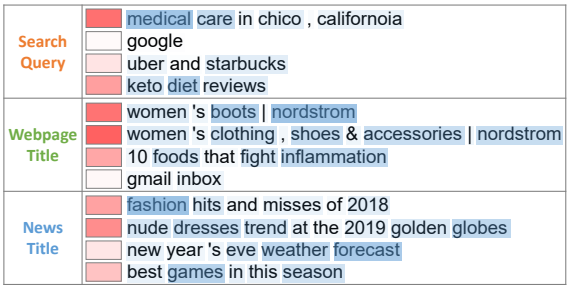

Figure 7: Visualization of the word-level and record-level attention weights from a randomly selected user.

tention networks in our approach. First, we visualize the distributions of the view-level attention weights in our *NRHUB* approach, and the results are shown in Fig. 6. According to the results, we find the attention weights of the news encoder view can be relatively low. This may be because the news browsing histories of some users are less informative for modeling these users. In addition, we find the attention weight of the webpage encoder view is higher than the query encoder view. This may be because webpage titles are usually complete sentences, while search queries are usually pieces of words. Thus, webpage titles contain richer information of user interests than search queries, which are more informative for representing users.

Next, we illustrate the attention weights in the word-level and record (news, query and webpage)-level attention networks. The results are shown in Fig. 7. From Fig. 7, we have several observations. First, we find search queries and webpage titles can provide much information of user interests which is not covered by the news browsing histories of users. For example, we can infer that this user is interested in the health topic from her queries and browsed webpages, which is not revealed by her clicked news. Thus, incorporating heterogeneous behaviors of users are useful for learning more accurate user representations. Second, we find that our approach can effectively select important words in queries, news and webpage titles. For example, the word "medical" gains high attention weight since it is informative for learning query representations, while the word "in" receives low attentions. In addition, our approach can effectively select important queries, news and webpages. For example, the query "deto diet reviews" and the webpage of "10 foods that fight inflammation" are assigned high attention weights since they are informative for learning user representations, but the query "google" and the webpage of "gmail inbox" are assigned low attention weights since they are uninformative. These results validate the effectiveness of the attention networks in our approach.

5 Conclusion and Furture Work

In this paper, we propose a neural news recommendation approach which can exploit heterogeneous user behaviors. In our approach, we propose an attentive multi-view learning framework to learn unified user representations from their search queries, clicked news, and browsed webpages by regarding them as different views. We propose to apply a view-level attention network to select important views for informative user representation learning. Besides, we propose to employ both word- and record-level attention networks to learn more informative news and user representations. Extensive experiments on a real-world dataset collected from MSN News show our approach can effectively improve the performance of neural news recommendation, and can achieve satisfactory results in cold-start scenarios.

In our future work, we will explore three potential directions. The first one is how to incorporate the interactions between different kinds of user behaviors, since the relatedness of them may be useful for modeling the interest evolution of users. The second one is how to incorporate the activities of users on social media platforms, since the opinions, experiences and events shared by users on social media are usually useful clues to indicate their interests. The third one is how to incorporate language models to generate context-aware word embeddings, which may enhance the representation learning of contexts in news and other kinds of user generated content.

Acknowledgments

The authors would like to thank Microsoft News for providing technical support and data in the experiments, and Jiun-Hung Chen (Microsoft News) and Ying Qiao (Microsoft News) for their support and discussions. This work was supported by the National Key Research and Development Program of China under Grant No. 2018YFC1604002, the National Natural Science Foundation of China under Grant Nos. U1836204, U1705261, U1636113, U1536201, and U1536207.

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. ACM.

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. ACM.

Abhinandan S Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. 2007. Google news personalization: scalable online collaborative filtering. In *WWW*, pages 271–280. ACM.

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

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. ACM.

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

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

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*.

Vaibhav Kumar, Dhruv Khattar, Shashank Gupta, and Vasudeva Varma. 2017. Word semantics based 3-d convolutional neural networks for news recommendation. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 761–764. IEEE.

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.

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

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

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Owen Phelan, Kevin McCarthy, Mike Bennett, and Barry Smyth. 2011. Terms of a feather: Content-based news recommendation and discovery using twitter. In *ECIR*, pages 448–459. Springer.

Steffen Rendle. 2012. Factorization machines with libfm. *TIST*, 3(3):57.

Jeong-Woo Son, A Kim, Seong-Bae Park, et al. 2013. A location-based news article recommendation with explicit localized semantic analysis. In *SIGIR*, pages 293–302. ACM.

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. 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. ACM.

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, Junxin Liu, Shaojian He, Yongfeng Huang, and Xing Xie. Neural demographic prediction using search query.

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 Li Guo. 2019. Dan: Deep attention neural network for news recommendation. In *AAAI*, volume 33, pages 5973–5980.