ISLF: Interest Shift and Latent Factors Combination Model for Session-based Recommendation

Jing Song[1] , **Hong Shen**[1,2] , **Zijing Ou**[1] , **Junyi Zhang**[1] , **Teng Xiao**[1] and **Shangsong Liang**[1]
[1]School of Data and Computer Science, Sun Yat-sen University, China
[2]School of Computer Science, The University of Adelaide, Adelaide, Australia
songj35@mail2.sysu.edu.cn, hongsh01@gmail.com, {ouzj,zhangjy329}@mail2.sysu.edu.cn,
{cstengxiao,liangshangsong}@gmail.com

Abstract

Session-based recommendation is a challenging problem due to the inherent uncertainty of user behavior and the limited historical click information. Latent factors and the complex dependencies within the user's current session have an important impact on the user's main intention, but the existing methods do not explicitly consider this point. In this paper, we propose a novel model, Interest Shift and Latent Factors Combination Model (ISLF), which can capture the user's main intention by taking into account the user's interest shift (i.e. long-term and short-term interest) and latent factors simultaneously. In addition, we experimentally give an explicit explanation of this combination in our ISLF. Our experimental results on three benchmark datasets show that our model achieves state-of-the-art performance on all test datasets.

1 Introduction

As a significant constituent part of modern electronic commerce systems, Session-based Recommendation System (SRS) is a subtask of Recommendation System. The task of SRS is to predict which item the user will click next based on the implicit feedbacks of the clicked item sequence in the current session [Hidasi *et al.*, 2015; Zhu *et al.*, 2017]. When a user starts to click an item, a session starts, the user will keep clicking on different items until the user's needs are met, the user stops clicking, and the session ends.

Current literatures [Hidasi *et al.*, 2015; Tan *et al.*, 2016; Li *et al.*, 2017; Liu *et al.*, 2018] have extensively investigated neural network based methods in SRS. Among them, recurrent neural network (RNN) [Hidasi *et al.*, 2015; Tan *et al.*, 2016; Li *et al.*, 2017] have received great attention, due to their capabilities in modeling the user's sequential behavior based on the user's clicked sequence. However, when a user accidentally clicks on inappropriate items or is attracted by irrelevant items due to curiosity, it is unreasonable to make recommendations only by modeling the user's sequential behavior. NARM [Li *et al.*, 2017] was proposed to model the user's sequential behavior and capture the user's main intention, although it tries to capture the main intention, it does not take into account the user's interest shift over time, which is also important for recommendation [Jannach *et al.*, 2015]. Liu et al. [Liu *et al.*, 2018] consider the user's interest shift by applying short-term and long-term interests to SRS. Regretfully, they cannot model the user's sequential behavior, which is very important because the order of the clicked items in the current session implies some useful information.

Although session-based recommendation has made great progress in past few years due to the application of various neural network models, it is still a challenging problem due to the inherent uncertainty of user behavior and the limited clicked information. To enhance recommendation performance, it is crucial to explore the factors that affect users' main intention and their future behavior. Inspired by the widespread application of latent variable models in collaborative filtering [Hofmann, 2004; Kabbur *et al.*, 2013; Wang and Blei, 2011], we believe that in SRS, they can still model the complex and hidden causal relationships that ultimately affect users' main intentions. Recently, in domains such as speech processing and computer vision, the paradigms based on the combination of probabilistic latent variable modeling and deep learning were proven quite effective [Kingma and Welling, 2013; Rezende *et al.*, 2014]. However, they have not yet been adopted in session-based recommendation.

In this paper, we consider the task of SRS from the perspective of the combination of neural network model and latent variable modeling and propose a model called Interest Shift and Latent Factors Combination Model (ISLF). Firstly, we introduce an improved variational autoencoder (VAE) to RNN at each timestamp to obtain the latent factor variables and the user's sequential behavior characteristics, and the last latent factor variable is used as the latent factor representation of the current session. Furthermore, in order to capture the user's long-term interest, we propose a novel attention mechanism to calculate the user's attention weight at each timestamp, then the weighted sum vector of all clicked item embeddings is used as the long-term interest representation. Especially, the attention weights are calculated based on the sequential behavior characteristics and the click frequency of each item in the current session. Since the user's short-term interest is the current interest that changes over time, We use the last clicked item embedding to represent it. Eventually, ISLF makes recommendations based on the user's interest shift (i.e. the long-term and short-term interest representations) and the latent factor representation. Our main contributions are as follows:

- We proposed ISLF model, which captures the complex and hidden causal relationships in the current session by introducing an improved variational autoencoder to RNN at each timestamp. It captures the user's main intention by considering the user's interest shift and the latent factors to make recommendations. To our knowledge, this is the first effort to introduce the combination of the variational autoencoder and neural network model into SRS to capture the complex and hidden causal relationships that will ultimately affect the user's main intention.
- A novel attention mechanism is proposed for the implementation of the ISLF model, in which the attention weights are calculated by combining the sequential behavior characteristics with the frequency of each item in the current session.
- Experimental results on two benchmark datasets show that ISLF achieves state-of-the-art, and the introduction of variational autoencoder and the proposed attention mechanism plays important roles.

2 Related Work

As a subtask of recommendation system, Session-based Recommendation System (SRS) is a typical application of recommendation system with implicit feedback. In SRS, users are considered to be anonymous, and their preferences (e.g. ratings) are not explicitly provided. On the contrary, decision makers can only refer to some implicit feedbacks from positive observations (e.g. click sequence of users in a series of times). In this section, we briefly review the related work on SRS from the following two aspects, i.e., general methods and sequence order based methods.

Normally, the general methods do not consider the order between items in the sequence. One approach is based on item-to-item recommendation, and the recommendations are made based on the similarity of items calculated according to the co-occurrence of items in sessions [Sarwar *et al.*, 2001; Linden *et al.*, 2003]. Recent approach, STAMP [Liu *et al.*, 2018] aims to capture users' long-term and short-term interests in order to obtain users' main intention. Although these methods are proved to be effective, they do not consider the sequential relationship between items, and thus user's sequential behavior implying useful information is ignored.

The sequence order based methods usually take into account the order within items of users' click sequences. The recommenders based on Markov chain utilizes sequential data information through the sequential clicked items of users' sequences [Gu *et al.*, 2014; Shani *et al.*, 2005; Zimdars *et al.*, 2001]. The main problem of the Markov chain based approaches is that when trying to include all possible selection sequences of users over all items, the state space quickly becomes unmanageable [Li *et al.*, 2017].

Deep neural networks have been used in SRS. Hidasi et al.[Hidasi *et al.*, 2015] is the first to apply recurrent neural networks (RNN) in SRS, by considering the user's clicked item sequence as the input of RNN. Tan et al.[Tan *et al.*, 2016] further improve the RNN-based model by using data augmentation technology and the means to explain the transfers in the input data distribution. Although the RNN-based methods mentioned above are obviously improved compared with the traditional recommendation methods, they only consider user's sequential behavior and do not emphasize user's main intention in the current session.

In SRS, besides user's sequential behavior, user's main intent is worthy of attention. Li et al. [Li *et al.*, 2017] proposed a model named NARM, which takes the last hidden state of RNN as a sequential behavior feature and calculate the attention weights of previously clicked item respectively, so as to capture the user's main intention. Although NARM have achieved better performance, we believe that there are many complex dependencies in sessions, and a good recommender should be able to simultaneously consider the user's interest shif, the user's sequential behavior and the latent factors in the current clicked session.

Unlike previous work, we propose a novel model, named ISLF to take into account the shift of user's interest and the latent factors simultaneously, and a novel attention mechanism to combine the sequential behavior characteristics and the frequency of each item in the current session with long-term interest. To our knowledge, this is the first effort to introduce variational autoencoder into session-based recommendation to capture the latent factors within the current session.

3 Methods

3.1 Symbolic Description

Let $\boldsymbol{V}{=}[v_1,v_2...,v_{|V|}]$ represents a set of unique items in the session-based recommendation system, called item dictionary. Each session is denoted by $\boldsymbol{S} = [s_1, s_2..., s_N]$, where $s_i(1 \leq i \leq N)$ is the index of the item clicked at i-th timestamp in the item dictionary. $\boldsymbol{S_T} = [s_1, s_2, ..., s_T], 1 \leq T \leq N$ denotes the prefix of $\boldsymbol{S}$ truncated at T-th timestamp. Given session prefix $\boldsymbol{S_T}$, the session-based recommendation task is to predict which item the user most likely to click next. In this paper, we proposed an Interest Shift and Latent Factors Combination Model (ISLF) to complete this task. Let $\boldsymbol{E} = [\boldsymbol{e_1}, \boldsymbol{e_2}, ..., \boldsymbol{e_{|V|}}]$ denote the item embedding representation corresponding to the item dictionary $\boldsymbol{V}$. Then for any given session prefix $\boldsymbol{S_T}$, let the corresponding item embedding representation $\boldsymbol{X_T} = [\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_T}]$ be the input of ISLF, where $\boldsymbol{x_i}$ denotes the item embedding vector of the clicked item s_i. Then we obtain the output $\hat{\boldsymbol{Y}}_{\boldsymbol{T}} = \mathrm{ISLF}(\boldsymbol{X_T})$, where $\hat{\boldsymbol{Y}}_{\boldsymbol{T}} = [\hat{y}_{T1}, \hat{y}_{T2}, ..., \hat{y}_{T|V|}]$ and $\hat{y}_{Ti}$ represents the probability of item v_i being clicked at next timestamp. Finally the top-$k(1 \leq k \leq |V|)$ items in $\hat{\boldsymbol{Y}}_{\boldsymbol{T}}$ are recommended.

3.2 The Recurrent VAE

As show in Figure 1, the input of ISLF is the click sequence $\boldsymbol{X_T} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_T}]$, then the input $\boldsymbol{X_T}$ is converted into sequential behavior characteristics $\boldsymbol{U_T} = [\boldsymbol{u_1}, \boldsymbol{u_2}, \cdots, \boldsymbol{u_T}]$ and latent factor variables $\boldsymbol{Z_T} = [\boldsymbol{z_1}, \boldsymbol{z_2}, \cdots, \boldsymbol{z_T}]$ through a recurrent VAE, which is implemented by adding a VAE to

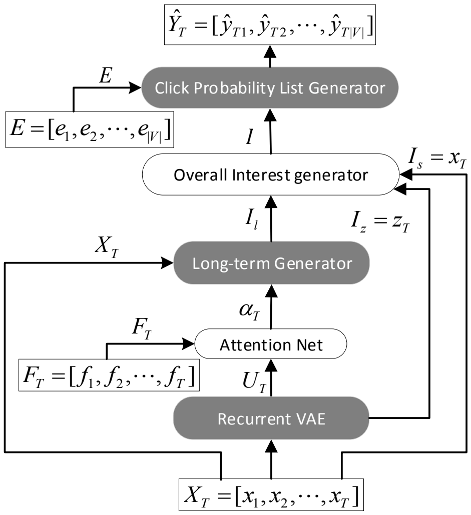

$\hat{Y}_T = [\hat{y}_{T1}, \hat{y}_{T2}, \cdots, \hat{y}_{T|V|}]$
E
Click Probability List Generator
$E = [e_1, e_2, \cdots, e_{|V|}]$
I
Overall Interest generator
$I_s = x_T$
I_l
$I_z = z_T$
X_T
Long-term Generator
α_T
F_T
Attention Net
$F_T = [f_1, f_2, \cdots, f_T]$
U_T
Recurrent VAE
$X_T = [x_1, x_2, \cdots, x_T]$

Figure 1: The general framework and dataflow of the proposed model ISLF.

RNN at each timestamp. The prior on the latent factor variable $\boldsymbol{z_t}$ follows the distribution:

$$\boldsymbol{z_t} \sim \mathcal{N}(\boldsymbol{\mu_{0,t}}, \mathrm{diag}(\boldsymbol{\sigma^2_{0,t}})), \quad (1)$$

where $\boldsymbol{\mu_{0,t}}$ and $\boldsymbol{\sigma_{0,t}}$ denote the parameters of the conditional prior distribution, which are obtained through a MLP cell:

$$[\boldsymbol{\mu_{0,t}}, \boldsymbol{\sigma_{0,t}}] = \varphi(\boldsymbol{W_0 h_t} + \boldsymbol{b_0}), \quad (2)$$

where $\varphi(\cdot)$ is a sigmoid function. Then the RNN updates the hidden states using the recurrence equation:

$$\boldsymbol{h_{t+1}} = g(\boldsymbol{x_{t+1}}, \boldsymbol{z_t}, \boldsymbol{h_t}), \quad (3)$$

where g is the Gated Recurrent Units. The approximate posterior on $\boldsymbol{z_t}$ will not only be a function of $\boldsymbol{h_t}$ but also of $\boldsymbol{x_{t+1}}$:

$$\boldsymbol{z_t} \mid \boldsymbol{x_{t+1}} \sim \mathcal{N}(\boldsymbol{\mu_{1,t}}, \mathrm{diag}(\boldsymbol{\sigma^2_{1,t}})), \quad (4)$$

similarly, $\boldsymbol{\mu_{1,t}}$ and $\boldsymbol{\sigma_{1,t}}$ denote the parameters of the approximate posterior distribution, which are obtained as follows:

$$[\boldsymbol{\mu_{1,t}}, \boldsymbol{\sigma_{1,t}}] = \varphi(\boldsymbol{W_1 h_t} + \boldsymbol{W_2 x_{t+1}} + \boldsymbol{b_1}), \quad (5)$$

Then the sequential behavior characteristic $\boldsymbol{u_t} \in \boldsymbol{U_T}$ is calculated on the basis of the hidden state $\boldsymbol{h_t}$ and the latent factor variable $\boldsymbol{z_t}$:

$$\boldsymbol{u_t} \sim \mathcal{N}(\boldsymbol{\mu_{2,t}}, \mathrm{diag}(\boldsymbol{\sigma^2_{2,t}})), \quad (6)$$

where $\boldsymbol{\mu_{2,t}}$ and $\boldsymbol{\sigma_{2,t}}$ denote the parameters of the distribution, which are obtained as follows:

$$[\boldsymbol{\mu_{2,t}}, \boldsymbol{\sigma_{2,t}}] = \varphi(\boldsymbol{W_{u1} h_t} + \boldsymbol{W_{u2} z_t} + \boldsymbol{b_u}). \quad (7)$$

where $\boldsymbol{W_0}, \boldsymbol{W_1}, \boldsymbol{W_2}, \boldsymbol{W_{u1}}, \boldsymbol{W_{u2}} \in \mathbb{R}^{d \times d}$ are weighting matrixes, and $\boldsymbol{b_0}, \boldsymbol{b_1}, \boldsymbol{b_u} \in \mathbb{R}^d$ are bias vectors.

In addition, the latent factor variables can capture the complex and hidden causal relationships that ultimately affect the user's main intention and are affected by the previous latent factor variables, we believe that the last latent factor variable $\boldsymbol{z_T}$ can extract the valid information of all previous latent factor variables, so the latent factor representation $\boldsymbol{I_z}$ is represented by $\boldsymbol{z_T}$:

$$\boldsymbol{I_z} = \boldsymbol{z_T}. \quad (8)$$

3.3 Long-term Interest Generator

The user's long-term interest in the current session is generally not easy to change over time, so it is necessary to know the user's attention weight on each clicked item. On one hand, we think the user's sequential behavior characteristics $\boldsymbol{U_T}$ can reflect some useful information, on the other hand, we think the frequency of each clicked item in the current session reflects some information, which are denoted by $\boldsymbol{F_T} = [f_1, f_2, \cdots, f_T]$. Some items have often been frequently accessed, the reason for this phenomenon is that the user is usually indecisive in their decision-making process [Liu *et al.*, 2015]. Therefore, the attention mechanism considers the frequency of each item in the current session and the sequential behavior characteristics simultaneously, and the attention weight α_{iT} is computed as follows:

$$\alpha_{iT} = \boldsymbol{W_{\alpha 0}} \varphi(\boldsymbol{W_{\alpha 1} u_T} + \boldsymbol{W_{\alpha 2} u_i} + \boldsymbol{W_{\alpha 3}} f_i + \boldsymbol{b_\alpha}), \quad (9)$$

where $\boldsymbol{u_i} \in \mathbb{R}^d$ and f_i denote the sequential behavior characteristic at i-th timestamp and the frequency of the i-th clicked item respectively, $\boldsymbol{u_T} \in \mathbb{R}^d$ is the last sequential behavior characteristic, $\boldsymbol{W_{\alpha 0}}, \boldsymbol{W_{\alpha 3}} \in \mathbb{R}^{1 \times d}$ are weighting vectors, $\boldsymbol{W_{\alpha 1}}, \boldsymbol{W_{\alpha 2}} \in \mathbb{R}^{d \times d}$ are weighting matrices, $\boldsymbol{b_\alpha} \in \mathbb{R}^d$ are bias vectors. After obtaining the attention weights $\boldsymbol{\alpha_T}$, the long-term interest generator can adaptively select the important clicked items through $\boldsymbol{\alpha_T}$ to compute the user's long-term interest representation $\boldsymbol{I_l}$ as follows:

$$\boldsymbol{I_l} = \sum_{i=1}^{T} \alpha_{iT} \boldsymbol{x_i}, \quad (10)$$

where $\boldsymbol{x_i}$ is the item embedding of item s_i. Note that the proposed attention mechanism is distinctly different from related works, which consider the sequential behavior characteristics and the frequency of each clicked item simultaneously.

3.4 ISLF

Figure 2 shows the schematic illustration of ISLF, the latent factor representation $\boldsymbol{I_z}$ can capture the complex and hidden causal relationships in the current session, and the long-term interest representation $\boldsymbol{I_l}$ can capture the user's interest which does not change easily over time. Besides, the user's current interest is always changing over time, so the last item embedding $\boldsymbol{x_T}$ is used as the short-term interest representation $\boldsymbol{I_s}$:

$$\boldsymbol{I_s} = \boldsymbol{x_T}. \quad (11)$$

Then the overall interest generator combine $\boldsymbol{I_z}$, $\boldsymbol{I_l}$ with $\boldsymbol{I_s}$ through a simple MLP cell, which can extract the important features of the user's main intention to generate the overall interest representation $\boldsymbol{I}$:

$$\boldsymbol{I} = \varphi(\boldsymbol{W_{I1} I_l} + \boldsymbol{W_{I2} I_s} + \boldsymbol{W_{I3} I_z} + \boldsymbol{b_I}), \quad (12)$$

where $\boldsymbol{W_{I1}}, \boldsymbol{W_{I2}}, \boldsymbol{W_{I3}} \in \mathbb{R}^{d \times d}$ are weighting matrices, $\boldsymbol{b_I} \in \mathbb{R}^d$ is a bias vector. Finally, for a given candidate item embedding $\boldsymbol{e_i} \in \boldsymbol{E}$, the unnormalized similarity between them is defined as:

$$c_{Ti} = \boldsymbol{e_i} \cdot \boldsymbol{I}, \quad (13)$$

Then a softmax function process $\boldsymbol{C_T} = [c_{T1}, c_{T2}, ..., c_{T|V|}]$ to obtain the output $\boldsymbol{\hat{Y}_T}$:

$$\boldsymbol{\hat{Y}_T} = \mathrm{softmax}(\boldsymbol{C_T}), \quad (14)$$

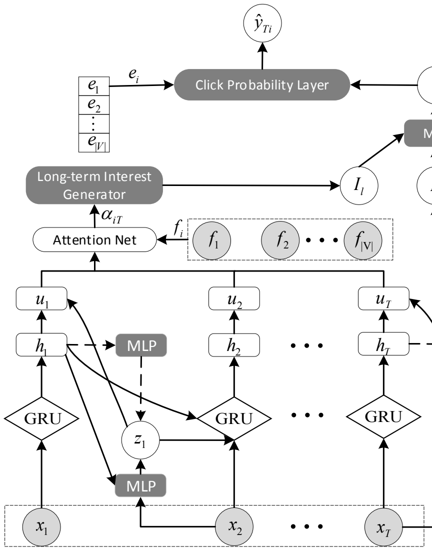


Figure 2: Schematic illustration of ISLF at T-th timestamp.

where $\hat{\boldsymbol{Y}}_{\boldsymbol{T}} = [\hat{y}_{T1}, \hat{y}_{T2}, ..., \hat{y}_{T|V|}]$ denotes the user's click probability list over the item dictionary $\boldsymbol{V}$ at next timestamp, each element $\hat{y}_{Ti} \in \hat{\boldsymbol{Y}}_{\boldsymbol{T}}$ denotes the probability of the event that item v_i is going to appear as the next-click in this session.

It is worth noting that, as can be seen from Eq. (3), $\boldsymbol{h_{t+1}}$ is a function of $\boldsymbol{x_{\leq t+1}}$ and $\boldsymbol{z_{\leq t}}$, therefore, Eq. (1) can define the distribution $P(\boldsymbol{z_t}|\boldsymbol{x_{\leq t}}, \boldsymbol{z_{<t}})$. Eq. (14) can define the distribution $P(\hat{\boldsymbol{Y}}_{\boldsymbol{T}}|\boldsymbol{x_{\leq T}}, \boldsymbol{z_{\leq T}})$, the parameterization of the generative model results in the factorization:

$$P(\hat{\boldsymbol{Y}}_{\boldsymbol{\leq T}}, \boldsymbol{z_{\leq T}}) = \prod_{t=1}^{T} P(\hat{\boldsymbol{Y}}_{\boldsymbol{t}}|\boldsymbol{x_{\leq t}}, \boldsymbol{z_{\leq t}}) P(\boldsymbol{z_t}|\boldsymbol{x_{\leq t}}, \boldsymbol{z_{<t}}). \quad (15)$$

Similarly, Eq. (4) can define the distribution $Q(\boldsymbol{z_t}|\boldsymbol{x_{\leq t+1}}, \boldsymbol{z_{<t}})$ and we also observe that this conditioning on $\boldsymbol{h_{t+1}}$ results in the factorization:

$$Q(\boldsymbol{z_{\leq T}}|\boldsymbol{x_{\leq T+1}}) = \prod_{t=1}^{T} Q(\boldsymbol{z_t}|\boldsymbol{x_{\leq t+1}}, \boldsymbol{z_{<t}}). \quad (16)$$

The objective function becomes a timestep-wise variational lower bound using Eq. (15) and Eq. (16):

$$\mathrm{E}_{Q(\boldsymbol{z_{\leq T}}|\boldsymbol{x_{\leq T+1}})} \Bigg[\sum_{t=1}^{T} (-\mathrm{KL}(Q(\boldsymbol{z_t}|\boldsymbol{x_{\leq t+1}}, \boldsymbol{z_{<t}})|| \\ P(\boldsymbol{z_t}|\boldsymbol{x_{\leq t}}, \boldsymbol{z_{<t}})) + \log P(\hat{\boldsymbol{Y}}_{\boldsymbol{t}}|\boldsymbol{x_{\leq t}}, \boldsymbol{z_{\leq t}})) \Bigg]. \quad (17)$$

So the loss function is as follows:

$$\mathcal{L}(\phi; X_T) = \frac{1}{2} \sum_{t=1}^{T} \Bigg[\sum_{k=1}^{d} (\log \sigma_{0,t,k}^2 - \log \sigma_{1,t,k}^2 + \frac{\sigma_{1,t,k}^2}{\sigma_{0,t,k}^2} - 1 \\ + \frac{(\mu_{0,t,k} - \mu_{1,t,k})^2}{\sigma_{0,t,k}^2}) - \sum_{i=1}^{|V|} (y_{Ti} \log(y_{\hat{T}i}) + (1 - y_{Ti}) \\ log(1 - y_{\hat{T}i})) \Bigg] \quad (18)$$

where ϕ denotes all the parameters of the model, d represents the item embedding dimension, $\sigma_{i,t,k}^2 (i = 0, 1)$ is the k-th element of $\mathrm{diag}(\boldsymbol{\sigma_{i,t}^2})$, $\mu_{i,t,k} (i = 0, 1)$ is the k-th element of $\boldsymbol{\mu_{i,t}}$, and $\boldsymbol{Y_T} = [y_{T1}, y_{T2}..., y_{T|V|}]$ denotes a one-hot vector exclusively activated by $s_{t+1} \in \boldsymbol{S}$ (the ground truth). For example, if s_{t+1} denotes the i-th element v_i in item dictionary $\boldsymbol{V}$, then $y_{Tk} = 1$, if $k = i$, and $y_{Tk} = 0$ if $k \neq i$. An iterative stochastic gradient descent (SGD) optimizer is then performed to optimize the cross-entropy loss.

4 Experiment

In this section, we detail our experimental setup.

4.1 Datasets and Data Preparation

We evaluate the proposed model and other models on the two datasets: (1) YOOCHOOSE, a public dataset released by RecSys'15 Challenge[1] and consists of click-streams gathered from an e-commerce web site. (2) DIGINETICA, a dataset from the CIKM Cup 2016[2], where we only use the transaction data. We filter out all sessions with length 1 and items with less than 5 occurrences in the datasets, and the items in the test set which do not appear in the training set. Same as [Liu *et al.*, 2018], we use a sequence splitting preprocess. That is, for the input session $\boldsymbol{S} = [s_1, ..., s_N]$, we generated the sequences and corresponding labels $([s_1], s_2)$, $\cdots$, $([s_1, s_2, \cdots, s_{n-1}], s_N)$ for both datasets, which proves to be effective. Because the training set of YOOCHOOSE is quite large and training on the recent fractions yields better results than training on the entire fractions as per the experiments of [Tan *et al.*, 2016], we use the recent fractions, 1/64 and 1/4, of training sequences, denoted as "YOO-1/64" and "YOO-1/4" datasets, respectively. The statistics of the three datasets (i.e., YOO-1/64, YOO-1/4 and DIGINETICA) are shown in Table 1.

4.2 Experimental Settings

Baselines. We compare ISLF to these baselines: (1) **POP** is a naive SRS which always recommends the most popular items in the training set; (2) **Item-KNN** [Davidson *et al.*, 2010] is an item-to-item model which recommends items similar to existing items, and the similarity is based on the co-occurrence number of two items in sessions; (3) **GRU-Rec** [Hidasi *et al.*, 2015] is a deep learning model which utilizes RNN with GRU units and session-parallel mini-batch training process. (4) **GRU-Rec+** [Tan *et al.*, 2016] is a improved model based on GRU-Rec. (5) **NARM** [Li *et al.*, 2017] is a model based RNN which employs attention mechanism to capture main purpose from the hidden states and takes the last hidden state as the sequential behavior feature, then combine the main purpose and the sequential behavior feature to generate recommendations; (6) **STAMP** [Liu *et al.*, 2018] is a deep learning model which takes the last item embedding as the short-term interest and employs attention mechanism based on all clicked item embeddings to capture the long-term interest, then combine them to make recommendations.

[1] http://2015.recsyschallenge.com/challenge.html

[2] http://cikm2016.cs.iupui.edu/cikm-cup

Datasets	YOO-1/64	YOO-1/4	DIGINETICA
train	375,073	5,969,416	719,470
test	55,898	55,898	60,858
clicks	565,552	7,980,529	982,961
items	17,694	30,660	43,097
avg.len.	6.16	5.71	5.12

Table 1: Statistics of the experiment datasets.

	YOO-1/64		YOO-1/4		DIGINETICA	
	Recall	MRR	Recall	MRR	Recall	MRR
POP	6.71	1.65	1.33	0.30	0.91	0.23
Item-KNN	51.6	21.81	52.31	21.70	28.35	9.45
GRU-Rec	60.64	22.89	59.53	22.60	43.82	15.46
GRU-Rec+	67.84	29.00	69.11	29.22	57.95	24.93
NARM	68.32	28.76	69.73	29.23	**62.58**	27.35
STAMP	68.74	29.67	70.44	30.00	62.03	27.38
ISLF-s	63.74	24.81	64.14	25.52	58.95	22.87
ISLF-l	67.36	29.27	67.95	28.15	60.75	27.21
ISLF-z	68.44	28.62	68.82	29.35	60.96	27.5
ISLF	**69.32**	**33.58**	**71.02**	**32.98**	62.09	**27.74**

Table 2: The overall performance over three datasets.

Variant Methods. To evaluate the reasonability of the proposed ISLF model, we propose the following variant methods of ISLF: (1)ISLF-z excludes the latent factor representation; (2) ISLF-l excludes the long-term interest; (3) ISLF-s excludes the short-term interest; (4) ISLF- is a method in which the hidden states and the sequential behavior characteristics are independent of the latent factor variables; (5) ISLF-z- is a variant method of ISLF-z, in which the hidden states and the sequential behavior characteristics are independent of the latent factor variables; (6) ISLF-a excludes the attention mechanism and the average vector of all clicked item embedding is taken as the user's long-term interest; (7) ISLF-f is a method in which the frequencies of all clicked items in the session are not considered when calculating the attention weights; (8) ISLF-u is a method in which the user's sequential behavior characteristics are not considered and are replaced by all clicked item embeddings.

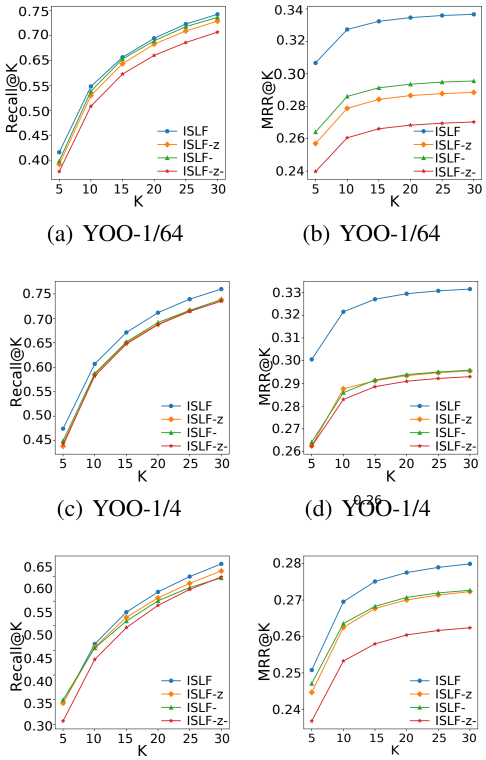
Evaluation Metrics. Recall@k and MRR@k (Mean Reciprocal Ran at k) are used to evaluate the performance of the SRS models. The Recall@k represents the proportion of test cases which have the desired items in top-k ranking lists in all test cases. The MRR@k represents the average of reciprocal ranks of the desire items. The reciprocal rank is set to zero if the rank is larger than k. In this paper, $k = 20$ is used for most tests.

Parameters Settings. The hyper-parameters are selected via extensive grid search on all the datasets, and the best models are optimized by early stopping based on the Recall@20 score on the validation set while using 10% of the training data as the validation set. The embedding dimension is searched in $\{50, 100, 200, 300\}$ and sets to 100 finally. The learning rate is searched in {0.001, 0.005, 0.01, 0.1, 1} and fixed at 0.005 finally. The mini-batch settings are: batch size:512, epoch:30. In Variational RNN, the number of MLP layer is selected in {1, 2, 3, 4, 5} and the optimal value is 4, the dim of mean and variance in VAE is the same as the embedding dimension, the coefficient of Kullback-Leibler divergence is selected in {0.001, 0.01, 0.1 , 1, 10} and finally set to 1, the GRU is set at 400 hidden units.

4.3 Experimental Results

Performance Comparison. In this part, we compare ISLF with the baselines and the variant methods (ISLF-z, ISLF-l and ISLF-s) to demonstrate the overall performance of ISLF as shown in Table 2. We have the following conclusions from Table 2: (1) The general method Item-KNN is not competitive, which proves the importance of the order within items of the clicked sessions. (2) All neural network based methods are superior to the general methods, which shows the effectiveness of deep learning technology in SRS. NARM and STAMP have achieved good performance in baselines, which indicates the importance of considering the user's main intention and the user's interest shift respectively. (3) In ISLF and its variants, ISLF-l has achieved performance comparable to GRU-Rec+, which may be due to the indirect consideration of the user's click order when considering latent variable factors, but it is still inferior to that of ISLF, which indicates the necessity of the user's long-term interest. The lower performance of ISLF-s also indicates the importance of taking the last click as the short-term interest. ISLF-z achieved best performance maybe since it takes into account the user's interest shift. Compared with ISLF-z, ISLF considers the latent factor representation in the current session, which achieves 1.26%, 3.10%, 1.85% improvements on Recall@20 and 17.33%, 12.37%, 0.87% on MRR@20 on three benchmark datasets respectively.

Effects of the Latent Factors. In this section, we design a series of comparison models to verify the validity of the latent factors obtained through the recurrent VAE. Figure 3 shows the experimental results of the Recall and MRR metrics on all datasets with recommendation list sizes ranging from 5 to 30. We can see that ISLF and ISLF- outperform ISLF-z and ISLF-z- respectively for all datasets, for example, for YOO-1/64, the ISLF averagely improves Recall and MRR by 2.70% and 17.32%, and ISLF- improves by 4.73% and 9.61%, which indicating the effectiveness of considering the latent factor representation when capturing the user's overall interest representation. We can also see that ISLF and ISLF-z outperform ISLF- and ISLF-z- respectively for all datasets, such as the ISLF averagely improves Recall and MRR by 1.40% and 14.40%, and ISLF-z improves by 3.41% and 6.88% for YOO-1/64, which proves that the latent factors are conducive to capturing the user's sequential behavior characteristics. In order to further demonstrate the performance of these models in different situations, we compare them on two groups of sessions of different lengths on YOOCHOOSE as Figure 4. Since the average length of sessions is almost 5, sessions with a length greater than 5 are called "long sessions" and the rest are called " short sessions". Compared with "long session", "short session" have higher performance on two metrics. Our guess is that as the session length increases, the user's interest continuously shifts and the main intention is difficult to capture. In addition, ISLF achieves the best in all situations, which just confirms the latent factors are beneficial to capture

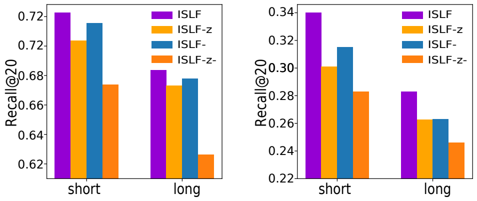


(a) YOO-1/64 (b) YOO-1/64

(c) YOO-1/4 (d) YOO-1/4

(e) DIGINETICA (f) DIGINETICA

Figure 3: (a) (c) (e) Recall and (b) (d) (f) MRR on all datasets for top-k recommendation.



(a) Recall (b) MRR

Figure 4: (a) Recall@20 and (b) MRR@20 of different session lengths on YOOCHOOSE.

the user's main intention in both long and short sessions.

	YOO-1/64		YOO-1/4		DIGINETICA	
	Recall	MRR	Recall	MRR	Recall	MRR
ISLF-a	68.43	32.52	70.04	30.30	60.48	26.53
ISLF-u	68.98	30.06	70.49	30.40	60.83	26.89
ISLF-f	68.80	29.32	70.35	29.63	61.58	26.94
ISLF	**69.32**	**33.58**	**71.02**	**32.98**	**62.09**	**27.74**

Table 3: Effects of the attention weights.

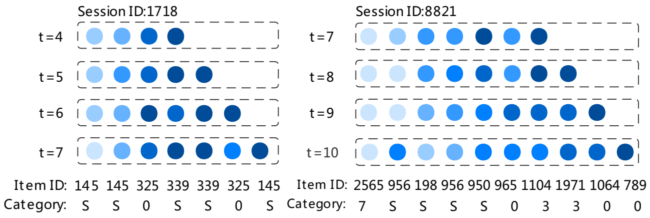


Figure 5: Visualization of attention weights.

The Proposed Attention Mechanism. In order to prove the effectiveness of the proposed attention mechanism, we designed a series of comparative experiments. From the results shown in Table 3, the following points can be observed: (1) The worst performance of ISLF-a indicates that it is difficult to judge which items users focuses on when excluding the attention mechanism. (2) The performance of ISLF-u is not as good as that of ISLF, which indicates that the click order is indeed conducive to capturing the main intention. (3) The performance of ISLF-f on Recall@20 and MRR@20 is 0.85% and 9.60% lower than that of ISLF on average, which proves that the items repeatedly clicked can also indicate the user's main intention to a certain extent. In order to intuitively illustrate the advantages of the attention mechanism, we visualized the attention weights in Figure 5, and for each item, the darker the color, the more attention the user pays to. From Figure 5, we can analyze the following points: (1) For the same item in the same session, the importance level is different at each timestamp, which indicates that the user's interest is constantly shifting over time, and our attention mechanism can select important items at different timestamps. (2) The category of item with high frequency in the current session is closely related to the user's main intention, which proves that the frequency of item in the current session does contribute to capturing the user's main intention. (3) No matter where an important item is in the session, it can always be captured, which shows that our attention mechanism can capture the important item in all directions to capture the user's main intention more accurately.

5 Conclusion

We propose ISLF model for the task of SRS and two important findings can be drawn from this study: (1) As user's clicked sequence contains many latent factors that affect the user's next click, ISLF aims at inferring its latent factors through the recurrent VAE, which can capture user's main intention more effectively. (2) The proposed attention mechanism can effectively capture user's long-term interest by combining the sequential behavior characteristics and the frequency of each item in the current session to estimate the shift of user interest. Experimental results show that ISLF model can achieve state-of-the-art performance on all testing datasets. As to future work, we intend to add auxiliary information such as attributes of users and items into our ISLF for performance improvement.

Acknowledgements

This work is supported by National Key R&D Program of China Project #2017YFB0203201 and Australian Research Council Discovery Project DP150104871. The corresponding author is Hong Shen.

References

[Davidson *et al.*, 2010] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, and Dasarathi Sampath. The youtube video recommendation system. In *Proceedings of RecSys 2010, Barcelona, Spain, September 26-30, 2010*, pages 293–296, 2010.

[Gu *et al.*, 2014] Wanrong Gu, Shoubin Dong, and Zhizhao Zeng. Increasing recommended effectiveness with markov chains and purchase intervals. *Neural Computing and Applications*, 25(5):1153–1162, 2014.

[Hidasi *et al.*, 2015] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. *CoRR*, abs/1511.06939, 2015.

[Hofmann, 2004] Thomas Hofmann. Latent semantic models for collaborative filtering. *ACM Trans. Inf. Syst.*, 22(1):89–115, 2004.

[Jannach *et al.*, 2015] Dietmar Jannach, Lukas Lerche, and Michael Jugovac. Adaptation and evaluation of recommendations for short-term shopping goals. In *Proceedings of the 9th ACM Conference on RecSys 2015, Vienna, Austria, September 16-20, 2015*, pages 211–218, 2015.

[Kabbur *et al.*, 2013] Santosh Kabbur, Xia Ning, and George Karypis. FISM: factored item similarity models for top-n recommender systems. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*, pages 659–667, 2013.

[Kingma and Welling, 2013] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[Li *et al.*, 2017] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. Neural attentive session-based recommendation. In *Proceedings of the 2017 ACM on CIKM 2017, Singapore, November 06 - 10, 2017*, pages 1419–1428, 2017.

[Linden *et al.*, 2003] Greg Linden, Brent Smith, and Jeremy York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, 2003.

[Liu *et al.*, 2015] Qi Liu, Xianyu Zeng, Chuanren Liu, Hengshu Zhu, Enhong Chen, Hui Xiong, and Xing Xie. Mining indecisiveness in customer behaviors. In *2015 IEEE International Conference on Data Mining, ICDM 2015, Atlantic City, NJ, USA, November 14-17, 2015*, pages 281–290, 2015.

[Liu *et al.*, 2018] Qiao Liu, Yifu Zeng, Refuoe Mokhosi, and Haibin Zhang. STAMP: short-term attention/memory priority model for session-based recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on KDD 2018, London, UK, August 19-23, 2018*, pages 1831–1839, 2018.

[Rezende *et al.*, 2014] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of ICML 2014, Beijing, China, 21-26 June 2014*, pages 1278–1286, 2014.

[Sarwar *et al.*, 2001] Badrul Munir Sarwar, George Karypis, Joseph A. Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of WWW 10, Hong Kong, China*, 2001.

[Shani *et al.*, 2005] Guy Shani, David Heckerman, and Ronen I. Brafman. An mdp-based recommender system. *Journal of Machine Learning Research*, 6:1265–1295, 2005.

[Tan *et al.*, 2016] Yong Kiam Tan, Xinxing Xu, and Yong Liu. Improved recurrent neural networks for session-based recommendations. In *Proceedings of the 1st Workshop on DLRS@RecSys 2016, Boston, MA, USA, September 15, 2016*, pages 17–22, 2016.

[Wang and Blei, 2011] Chong Wang and David M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011*, pages 448–456, 2011.

[Zhu *et al.*, 2017] Yu Zhu, Hao Li, Yikang Liao, Beidou Wang, Ziyu Guan, Haifeng Liu, and Deng Cai. What to do next: Modeling user behaviors by time-lstm. In *Proceedings of IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 3602–3608, 2017.

[Zimdars *et al.*, 2001] Andrew Zimdars, David Maxwell Chickering, and Christopher Meek. Using temporal data for making recommendations. In *UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, University of Washington, Seattle, Washington, USA, August 2-5, 2001*, pages 580–588, 2001.