From Past to Present: Personalized Attention Session-Aware RNN Recommender System

Presenter: Mei Wang
Mentors: Weizhi Li, Yan Yan
08/22/2018

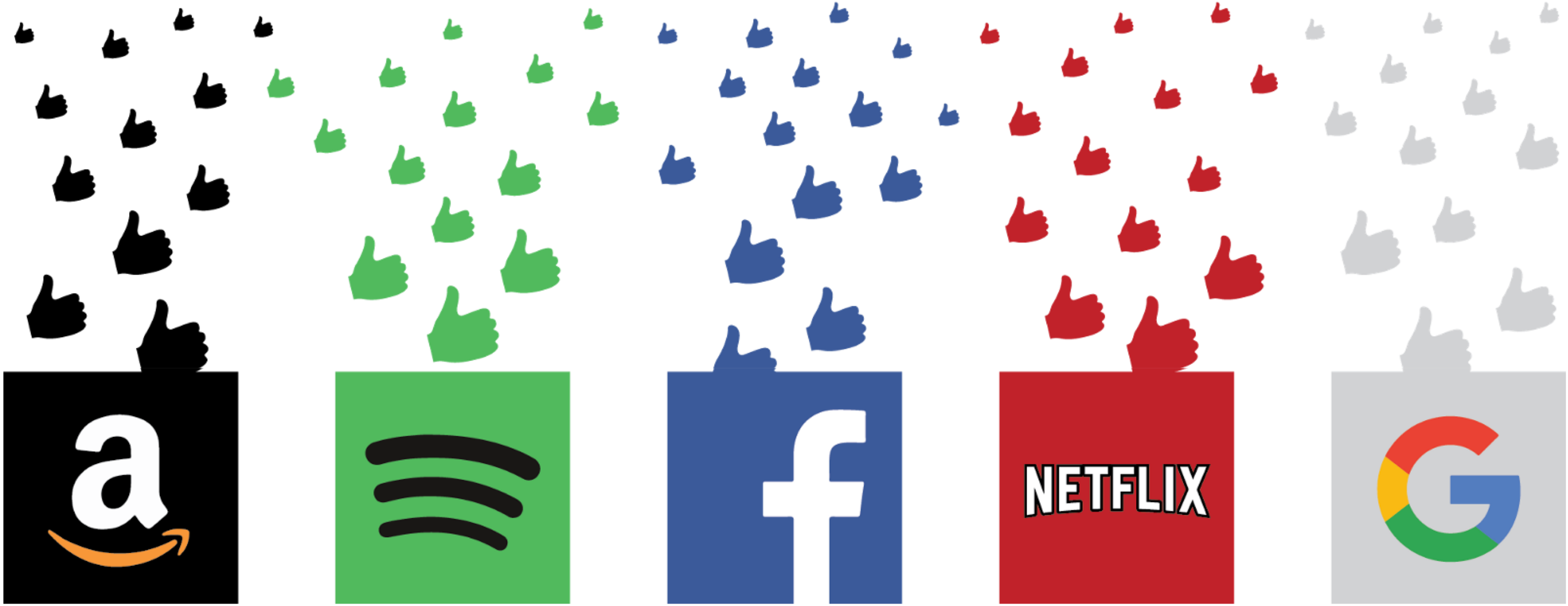
JD.COM

Outline

- 1. Introduction
- 2. Related Work
- 3. PASAR Model
- 4. Evaluation
- 5. Conclusion
- 6. Discussion

Introduction: Recommender Systems

↑↓Expanding ←→Wide-Ranging



e-commerce | music | social media | video | app store

Introduction: Personalized Recommendation

Influential Factors

Internal

Personality

Culture

Fashion Style

Aesthetic Taste

Age

Education

Figure

Marriage

...

External

Environments

Weather

Festivals

Location

Ads

Friends

Family

Income

...

Past

Periodic Purchases

Makeups

Car wiper blade

Favorite brands

Preferred Color

Clothing styles

Fast Moving Consumer Goods

...

Present

Current Needs

Umbrella

Birthday

Travel suits

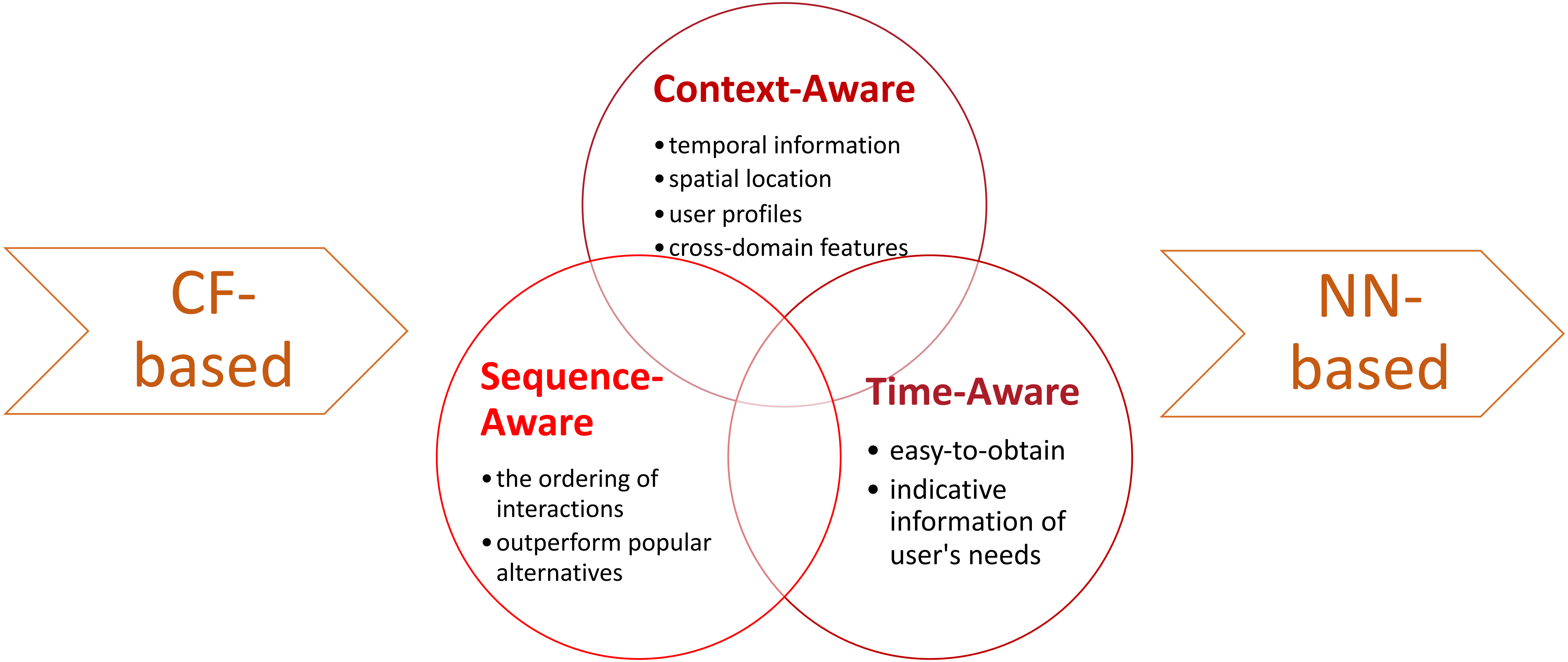
Mood

Promotions

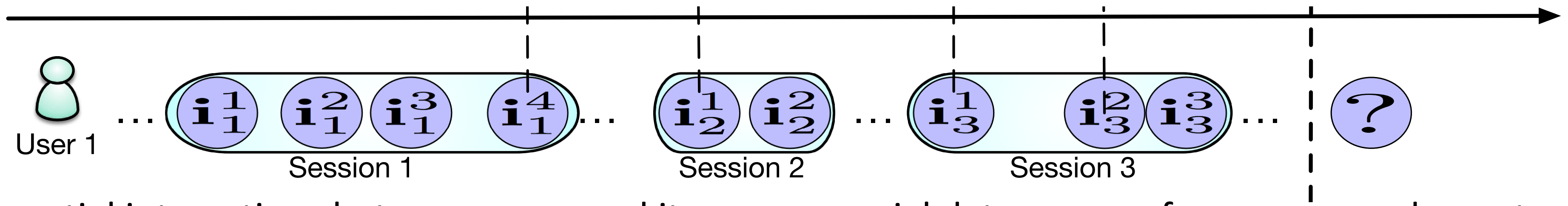
Hang around

...

Introduction: Rec Sys Development



Introduction: Limitation and Motivation



The sequential interactions between users and items are crucial data sources for recommender systems.

Limitation

Most existing session-based RNN methods solely focus on short-term user interactions within a single session and completely discard all the other long-term user interaction data cross different sessions.

Motivation

The goal of this work is to make effective use of both intra-session and inter-session profiles and construct a better personalized session-aware recommender system.

Introduction: Challenges

1

Traditional RNN cannot train with too long sequence length, which will result in extreme training latency and large memory cost.

2

The interaction data is very noisy: some clicks are meaningful, some are clicked by small interest, while some may even be clicked by mistake.

3

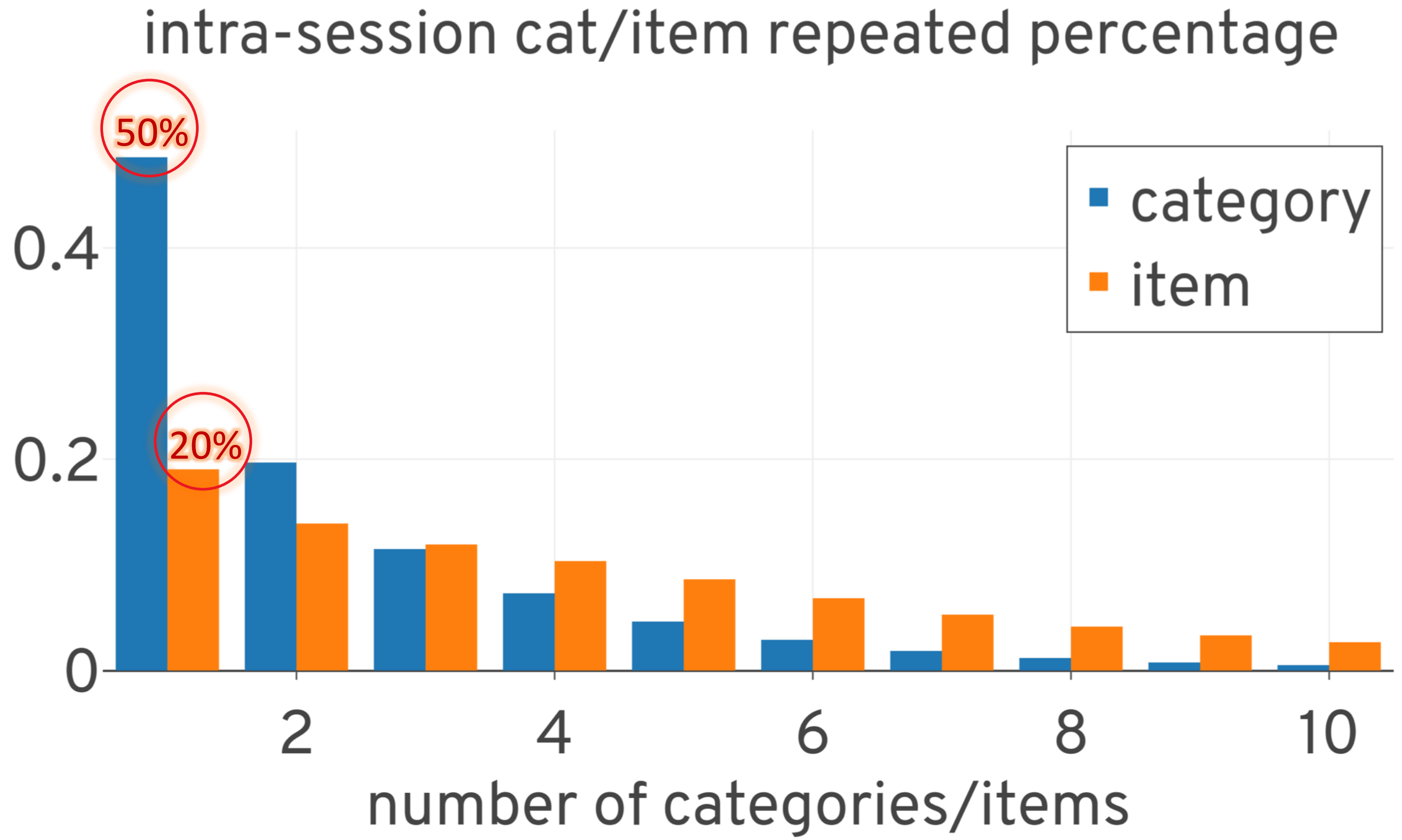
Data from the past sessions should play as different roles as present session, but there is no specific rule for integrating the session–based short–term profiles and session–aware long–term profiles.

Introduction: Motivated by Empirical Data Analysis

- The motivation of our model design is inspired by real data observations and analysis.
- It comes from online Tianchi e-commerce navigation log data having around 100M interactions, 1M users and 4M items from 10K categories.
- We come up with the following three observations.

Introduction: Empirical Data Analysis

1 Short-term profile predominates in the selection of the recommendations

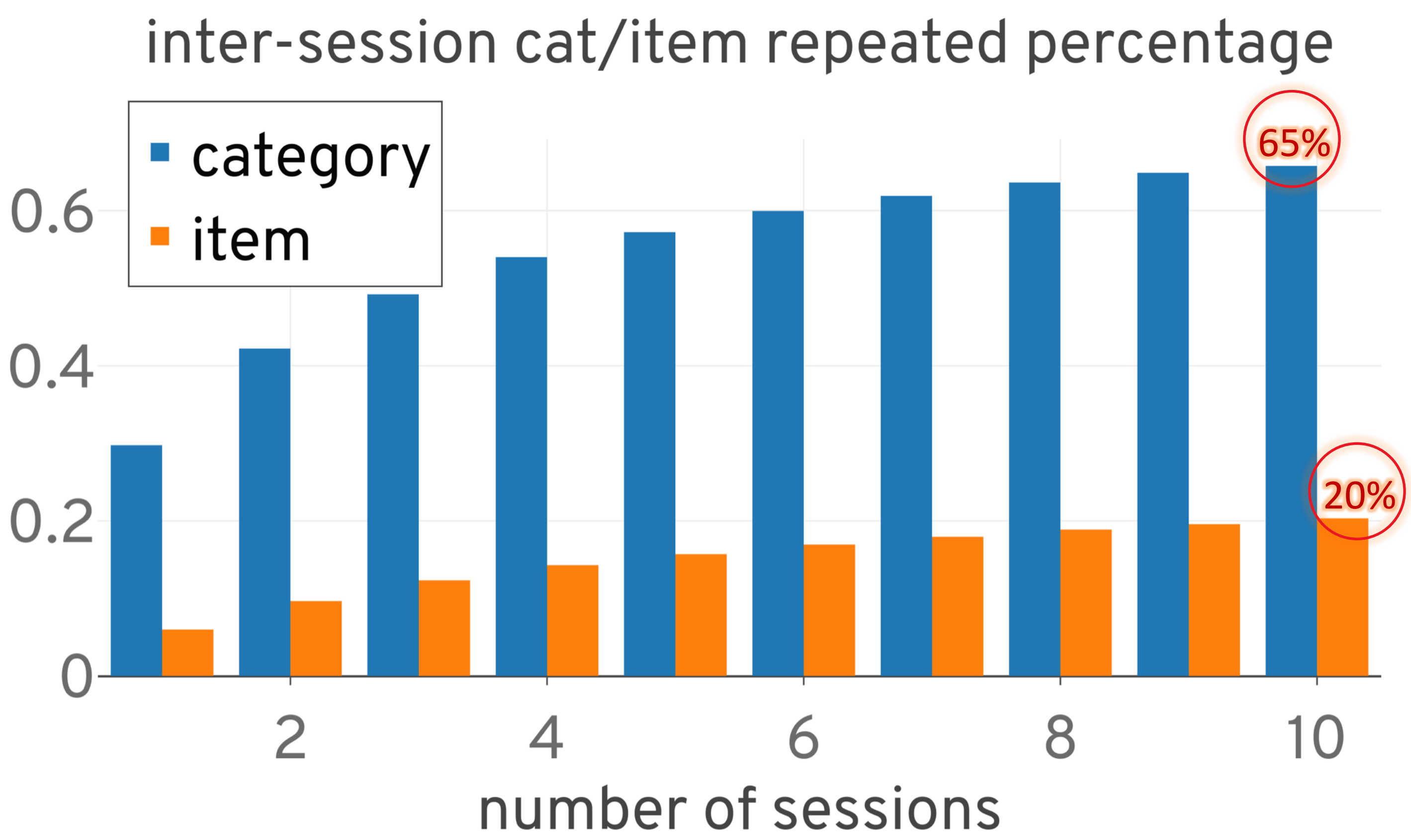


The blue bar represents the mean percentage of user interactions hanging in the top 10 categories during the same session, and the orange one represents for items.

Overall both of them are subject to exponential decrease, which proves that user's short-term shopping goal plays a predominant role for the intra-session interaction choices.

Introduction: Empirical Data Analysis

2 Longer-term behavioral patterns and user preferences can also be important

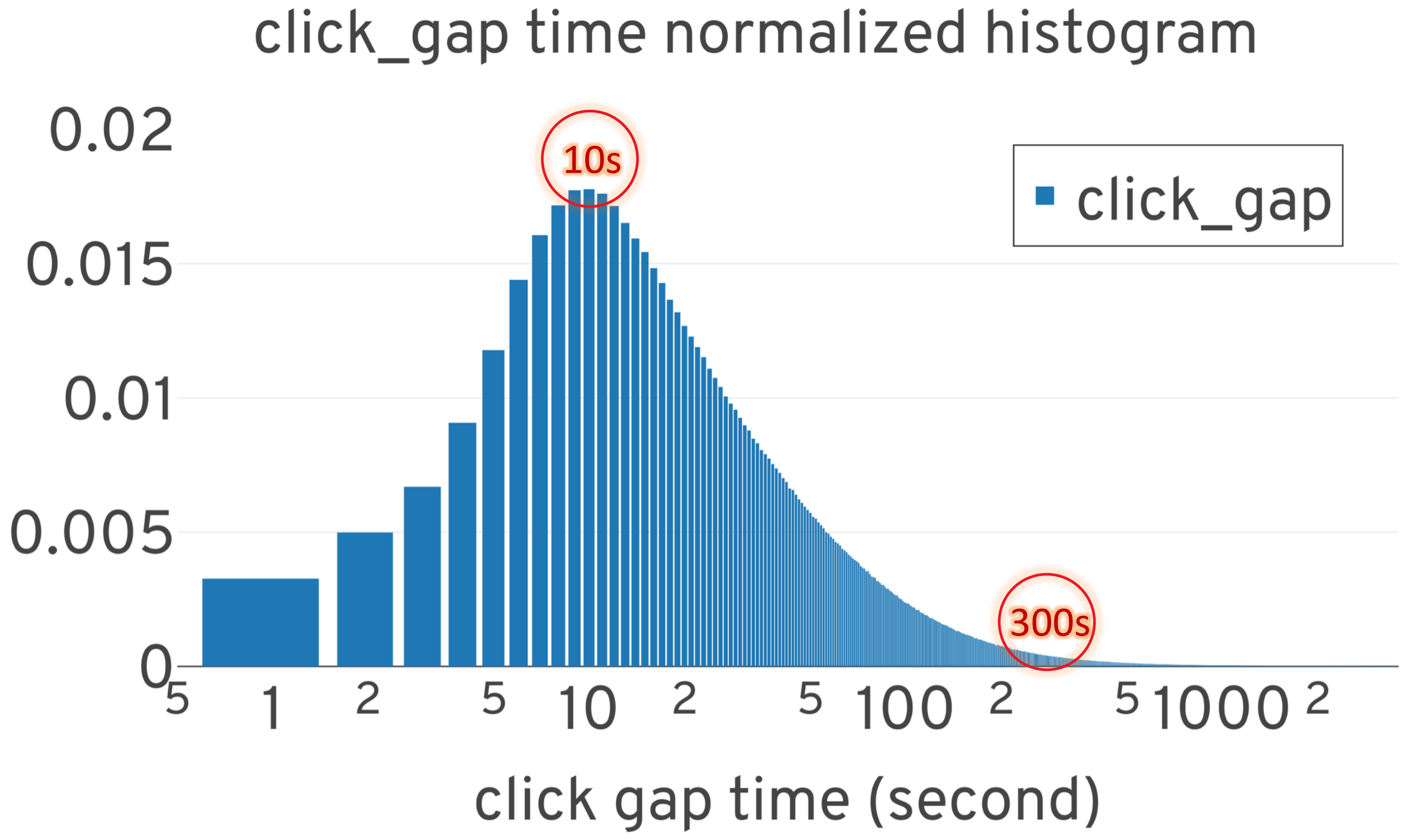


The figure shows mean percentage value of a user clicking some repeated categories and items that he/she had clicked before in the previous 10 sessions.

Inter-session information provides 30 to 60 percent of information for next-basket category prediction and 5 to 20 percent of knowledge about repeated items.

Introduction: Empirical Data Analysis

3 The click gap time (item dwell time) helps in connecting short-term and long-term



The figure shows the normalized histogram of the click gap (view dwell time) of user interactions, which follows gamma distribution with maximum around 10 seconds.

Generally speaking, the longer time a user spends on the item, the more interest he has in it. This perfectly bridges the gap of discrete interaction sequence data with potential weights.

Introduction: Motivated by Empirical Data Analysis

- In this paper, we want to quantify, exploit and integrate the effectiveness of user's intra-session and inter-session profiles with temporal dynamics.

1 Short-term profile predominates

The very last actions in the present session should represent an important piece of context information

GRU4REC

Session-based RNN recommender system as the basis of our model design

2 Longer-term profile counts

long-term profiles are important for recommender system, while current state-of-art session-based approaches fail to model them effectively.

We choose to use an efficient embedding layer to automatically train and activate short and long term profiles from session representations.

3 The dwell time helps

Finally, with the help of temporal dynamics scheme, we incorporate temporal context in the RNN and perform efficient combination for short-term session sequence information and long-term user and item profiles.

Introduction: Contributions

Personalized Session–Aware RS

We propose PASAR, a novel Personalized Attention Session–Aware Recommender system model, to seamlessly integrate intra–session and inter–session profiles.

Temporal Dynamics by Attention Net

We offer an extendable attention scheme to leverage temporal dynamics scheme exploiting more intra–session information so as to enhance session–based RS in time dimension.

Activate Long–term in Session RS

We include long–term user profiles for session–based RS to learn the cross–session pattern and user favorite evolution in a seamless way.

Empirical Results

We conduct extensive experiments on four real datasets and demonstrate the effectiveness of PASAR for personalized recommendation.

Related Work: Overview

Methods \ Features		General: User Taste	General: Item Impression	Multiplicative: User-Item Interaction	Evolution: User Favorite	Evolution: Item Trend	Time: Temporal Drift	Sequence: Sequence Feature
Notation		p_u	q_i	b_{ui}	$b_u(\mathrm{t})$	$q_i(t)$	t	seq
CF-based	BPR-MF	✓	✓	✓	×	×	×	×
	TimeSVD++	✓	✓	✓	✓	✓	✓	×
	FPMC	✓	✓	✓	×	×	×	✓
NN-based	DNN	✓	✓	✓	×	×	×	×
	GRU4REC	×	✓	✓	×	×	×	✓
	PASAR	✓	✓	✓	✓	✓	✓	✓

Related works compared by different methodology categories exploiting various domain features.

Related Work: CF−Based Rec Sys

2008 Netflix Prize

CF-Based Methods → **Context-Aware** → **Time-Aware** → **Sequence-Aware**

CF-Based Methods

- **Collaborative Filtering**
  (**CF**, Schafer et al, AdaptiveWeb’07)
- **Matrix Factorization**
  (**MF**, Koren et al, Computer’09)

Although CF-based methods have been theoretically well developed and are less expensive for computational cost, their practicableness and scalability yield to NN-based approaches.

Context-Aware

- **SVD++** model
  (Koren et al, KDD’08)
- **BPR-MF** model
  (Rendle et al, AUAI Press’09)
- **kNN** model
  (Koren, TKDD’10)
- Behavior Factorization
  (Zhao et al, WWW’17)

Time-Aware

- **TimeSVD++** model
  (Koren et al, KDD’09)

Sequence-Aware

- Factorizing Personalized Markov Chain
  (**FPMC**, Rendle et al, WWW’10)

Related Work: NN−Based Rec Sys

2016 YouTube DNN → **NN-Based Methods** → **Context-Aware** → **Time-Aware** → **Sequence-Aware**

NN-Based Methods

- **YouTube** DNN Model (Covington et al, RecSys'16)
- Neural Collaborative Filtering (**NCF**, He et al, WWW'17)

Context-Aware

- Deep & Cross model (**DCN**, Wang et al, ADKDD'17)
- **Latent Cross** model (Beutel et al, WSDM'18)

Time-Aware

- **Time-LSTM** model (Zhu et al, IJCAI'17)
- Recurrent Recommender Networks (**RRN**, Wu et al, WSDM'17)
- Neural Survival Recommendation (**NSR**, Jing et al, WSDM'17)
- Temporal Point Process (T**PS**, Song et al, GaTech)

Sequence-Aware

- Session-based RNN (**GRU4REC**, Hidasi , ICLR'16)

While these approaches did not adapt to session-based scheme, which could play a predominant actor for recommendation as shown in EDA.

Related Work: Session–Based Rec Sys

2016 GRU4REC RNN

Session-Based Methods

- **GRU4REC RNN Model (Hidasi et al, ICLR'16)**

Improvements

- **Data augmentation (Improved RNN, Tan et al, DLRS'16)**
- **Exploit dwell time (Dallmann et al, RecSys'17)**
- **Add different types of interactions (Wu et al, CIKM'17)**
- **Hierarchical RNN (Quadrana et al, RecSys'17)**

These works made incremental improvements for GRU4REC, but they do not make significant modification and haven't consider long-term intra-session info and user action gap time feature, which can make great gain.

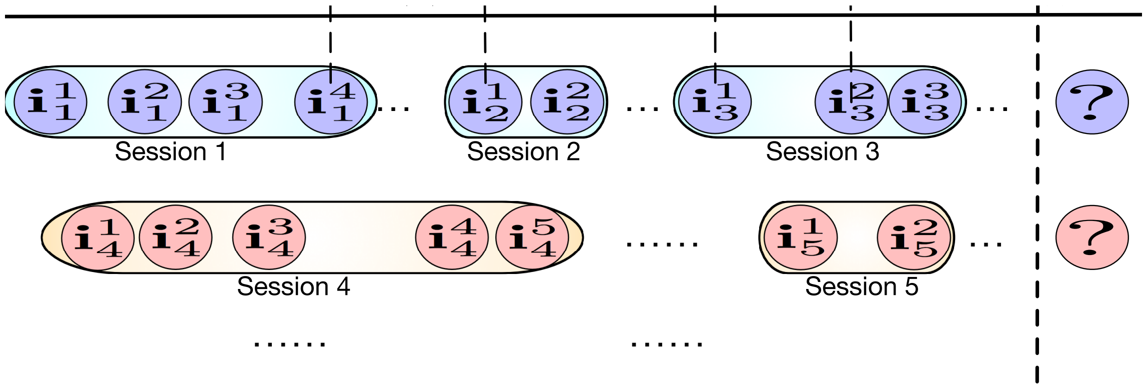
Related Work: Long and Short-term Combination

	Long-term	Short-term
[CIKM'15] STAR	LDA (Latent Dirichlet Allocation)	MCMC (Monte Carlo Markov Chain)
[RecSys'17] HRNN	GRUuser	GRUses

To sum up, RNNs show their privilege in short-term sequential pattern mining than other item-based or Markov Chain-based approaches.

To facilitate RNN with long-term profiling, the goal of this paper is to make effective use of both long-term and short profiles and construct a better personalized session-aware RNN recommender system.

PASAR Model: Problem Formulation



Data flow of user and item interactions over time.

Definition of session:

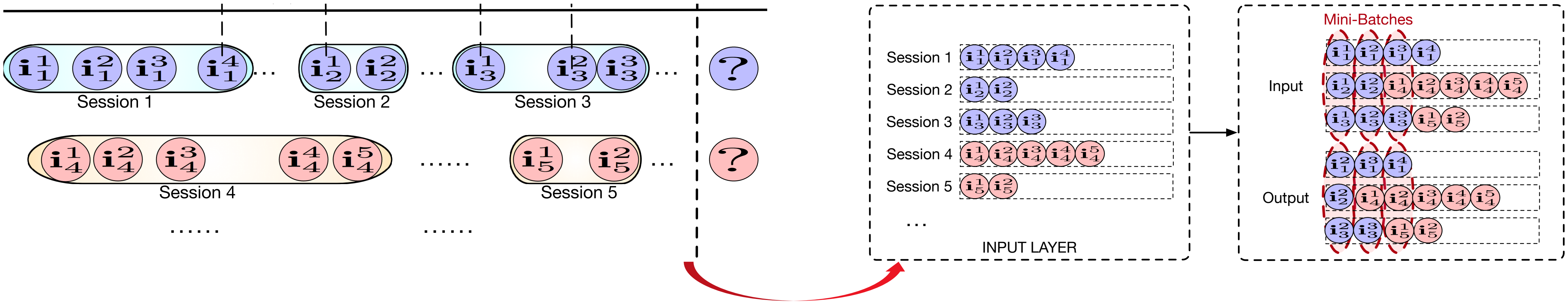
We define a session as a set of continuous navigation activities without interruption in the log sequence.
In our settings, we separate each session by at least one−hour inactivity.

We define the activity session sequence S as: $S = \{s_j \mid j = 1, \ldots, m\}$ Totally there are m sessions.

Where s_j represents the number j_{th} session: $s_j = \{i_j^1, i_j^2, \ldots, i_j^{n_j}\}$ Each session are of length n_j.

We formulate this as a top-K ranking problem: $\hat{r}_k = f(i^n \mid i^{1,\ldots,n-1})$ where $i \in I$ and I is the item set.

PASAR Model: Session−based RNN Framework



Different session length → session-parallel mini-batch approach $E = \{e_j\} = \{e_j^1, e_j^2, \dots, e_j^{n_j}\}$

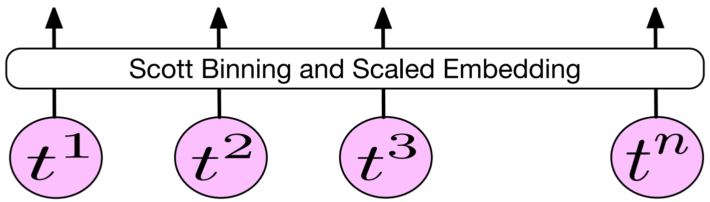
One-hot mini-batch vector is fed into a GRU layer, and the hidden states are reset when switching sessions.

The output of RNN can be treated as session-representations: $h_{session} = \mathrm{GRU}(e_j, h_{session-1})$

The likelihood in predictor is: $\hat{r}_k = g(e_k, h_k)$

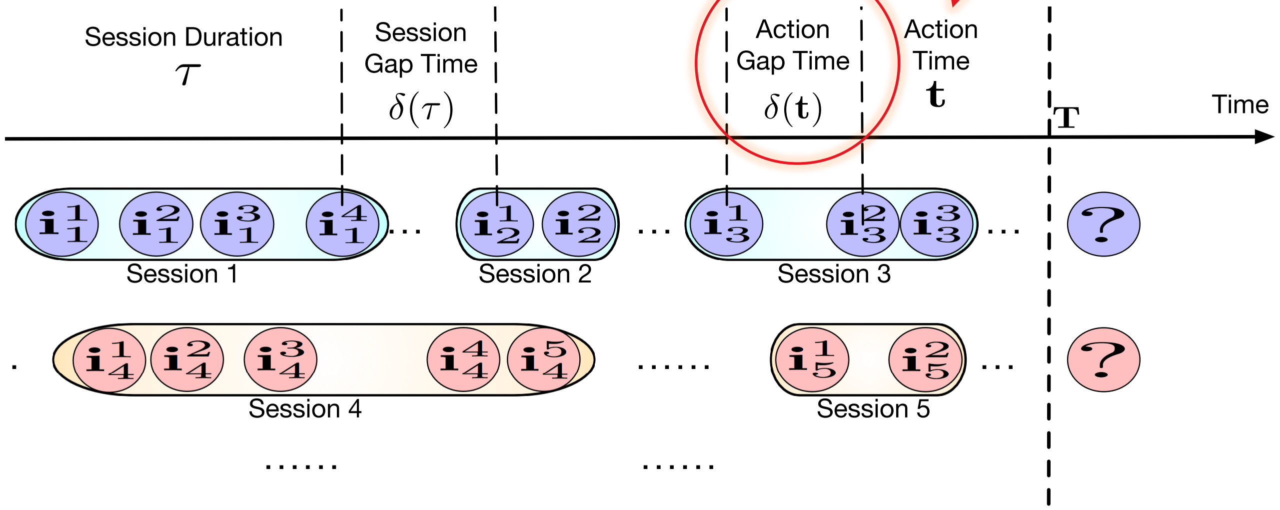
✓ Intra-session **Independent**

PASAR Model: Use of Item Dwell Time



Session gap time is helpful for survival analysis to predict user return time.

Action timestamp can be used for periodical purchasing feature training directly as contextual information.



Data flow of user and item interactions over time.

We create a dwell time sequence with the same dimension of item sequence:

$$t_j = \{t_j^1, t_j^2, \dots, t_j^{n_j}\}$$

Since it follows gamma distribution, we can take Scott binning of time to reduce dimensionality and accelerate the training process:

$$t_{bin} = \sigma \sqrt[3]{\frac{24 * \sqrt{\pi}}{n}}$$

Next, we use an embedding method to represent dwell time importance within sessions:

$$\text{E} = \{e_{t,j}\} = \{e_{t,j}^1, e_{t,j}^2, \dots, e_{t,j}^{n_{t,j}}\}$$

PASAR Model: Use of Item Dwell Time

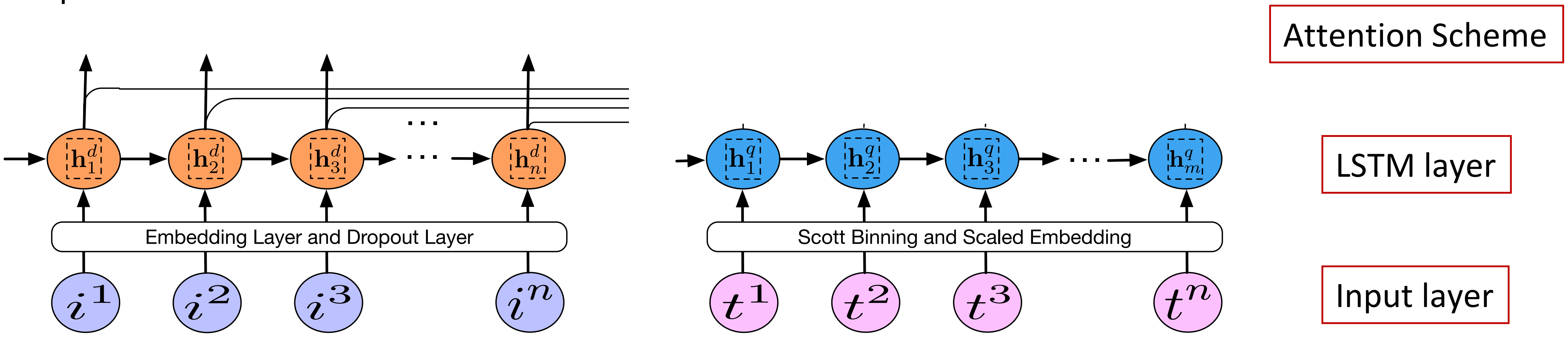
Attention Design:

Intuitively, such attention vectors are perfectly used to modulate the outputs of hidden states representing session orders, and it's reported as a very useful tool to extract the importance of sequence vector.

Attention Problems:

1. Sequence-In-Sequence-Out RNNs in NLP tasks

2. Enable data augmentation to get more training samples, all subsequences need to be forward to the attention network



PASAR Model: Use of Item Dwell Time

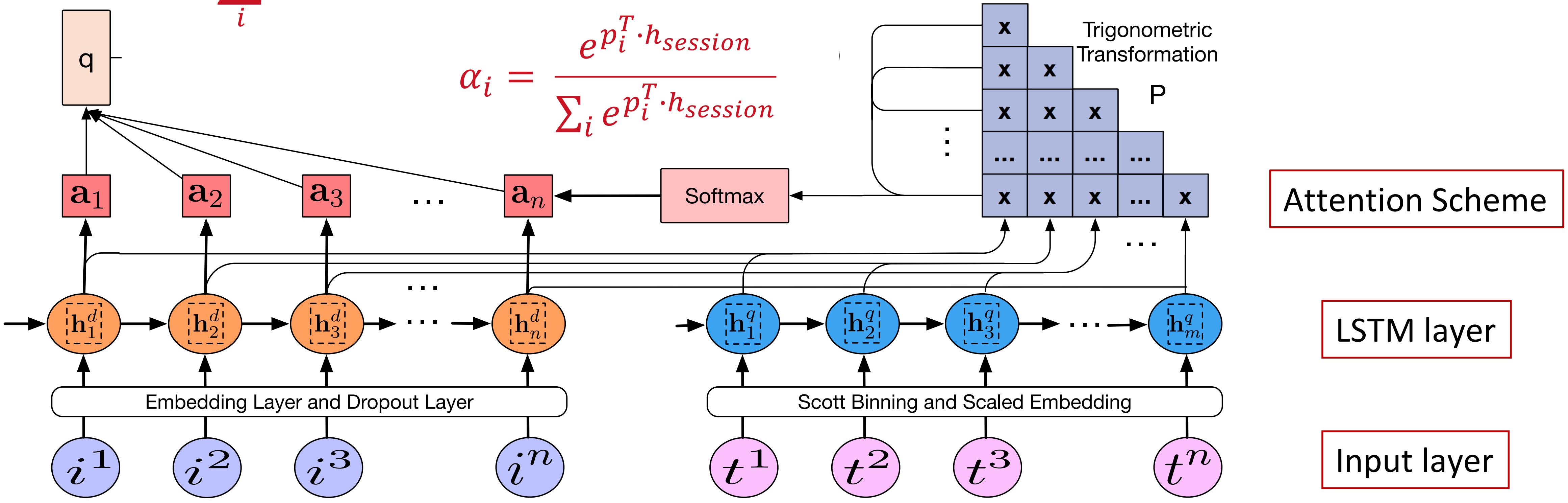
Triangle parallel attention method:

$$P = [P_i] = [p_0^T \cdot h_s, \dots, p_i^T \cdot h_s, 0, \dots, 0]$$

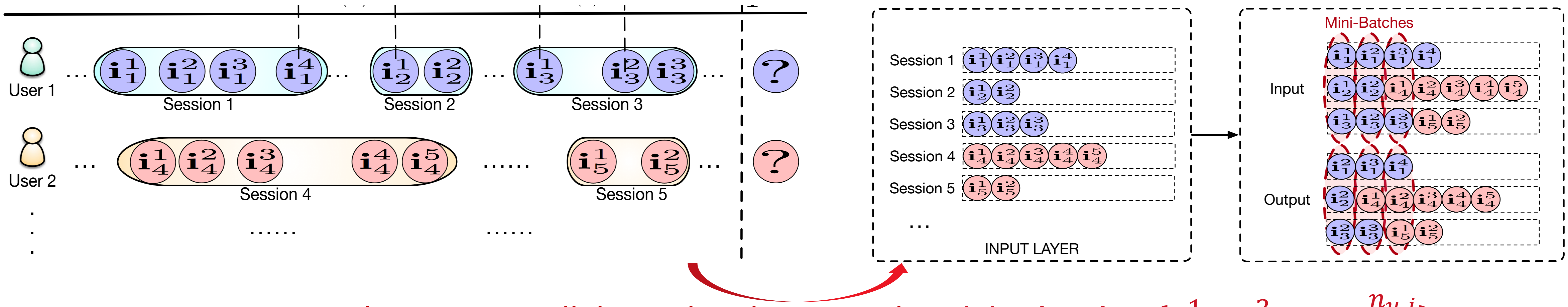
$$q_t = \sum_i \alpha_i \cdot h_{session}$$

$$p_i = \tanh(W_s h_{time} + b_s)$$

$$\alpha_i = \frac{e^{p_i^T \cdot h_{session}}}{\sum_i e^{p_i^T \cdot h_{session}}}$$



PASAR Model: Use of User Long−Term Profile



User-grouped session-parallel mini-batch approach: $E(u) = \{e_{u,j}\} = \{e_{u,j}^{1}, e_{u,j}^{2}, \ldots, e_{u,j}^{n_{u,j}}\}$

User-based negative sampling:

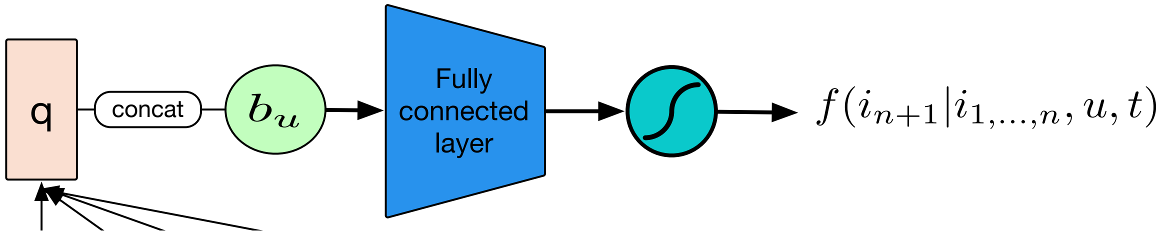
We select negative samples in proportion to the item popularity within mini-batch sequences.

Furthermore, for each user, we need to rule out the items appeared in his/her history.

This way, the local negative sampling method not only improves performance but also reduces the computational time.

PASAR Model: Use of User Long-Term Profile

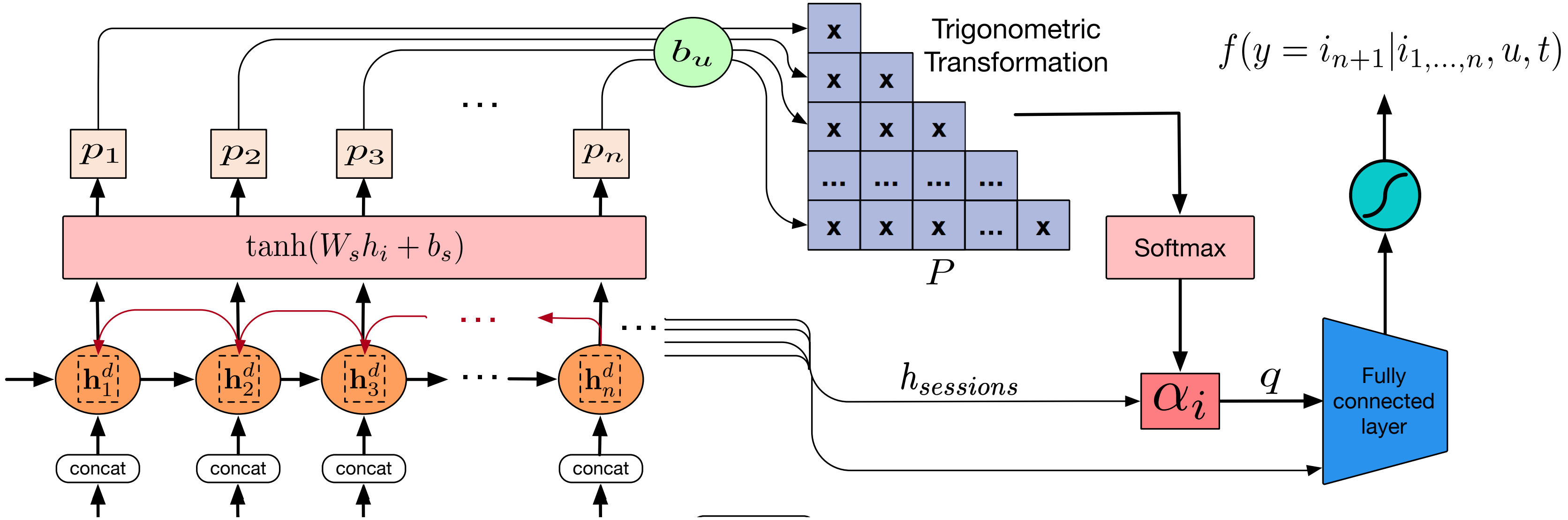
Concatenation Design:



The likelihood in predictor is:

$$\hat{r}_{j,k} = g\left(q_t \cdot e_k + b_j + b_k\right)$$

Attention Design:



Self-attention mechanism:

$$\alpha_i = \frac{e^{p_i^T \cdot e_u}}{\sum_i e^{p_i^T \cdot e_u}}$$

PASAR Model: Improving Extensions

Loss functions:

BPR loss: $$\mathrm{L} = -\frac{1}{N_s}\sum_{j=1}^{N_s}\log(\sigma(\hat{r}_j - r_k))$$

TOP1 loss: $$\mathrm{L} = \frac{1}{N_s}\sum_{j=1}^{N_s}\sigma\left(\hat{r}_j - r_k\right) + \sigma(\hat{r}_j{}^2)$$

Hinge loss: $$\mathrm{L} = \max\{\left(\hat{r}_j - r_k\right) + 1, 0\}$$

Data augmentation:

First, we train each sequence with all hidden outputs and make the predictions, which fully explores the subsequences information.

Second, we leverage the dropout layer for the sequences such that it makes regularization as well as diversifies the input sequence data.

Evaluation: Datasets

Totally we use four datasets in our experiments.

	user	time		both
Datasets	$MovieLens$	$Recsys15$	$Tianchi$	JD
Events	53,309	17,920,066	6,921,446	254,398
Users	237	/	12,332	3,035
Items	1,395	23,459	31,893	1,173
Sessions	3,609	4,247,567	93,287	45,878
Session Support	2	2	2	2
Item Support	10	20	10	20
User Support	10	/	10	20

Evaluation: Comparison Baselines and PASAR Versions

Models		Description
Baseline Models	BPR-MF	Matrix factorization techniques apply SVD factoring the user-item rating matrix
	YouTube DNN	YouTube model includes two stages: candidate generation and ranking
	WaveNet CNN	Inner multiplicative can be exploited by its stacked causal atrous convolutions
	GRU4REC RNN	Basic GRU layers and TOP-1 loss and session-parallel mini-batching mechanism
PASAR Variants	PASAR_user_att	Adding user profile embedding by self-attention network
	PASAR_user_cat	Adding the user profile by concatenating hidden outputs and user embeddings
	PASAR_time_att	Adding time profile embedding by global attention network
	PASAR_time_cat	Adding time profile by concatenating time embeddings and item embeddings
	PASAR_time_user	Integrating both time and user profiles as final PASAR model

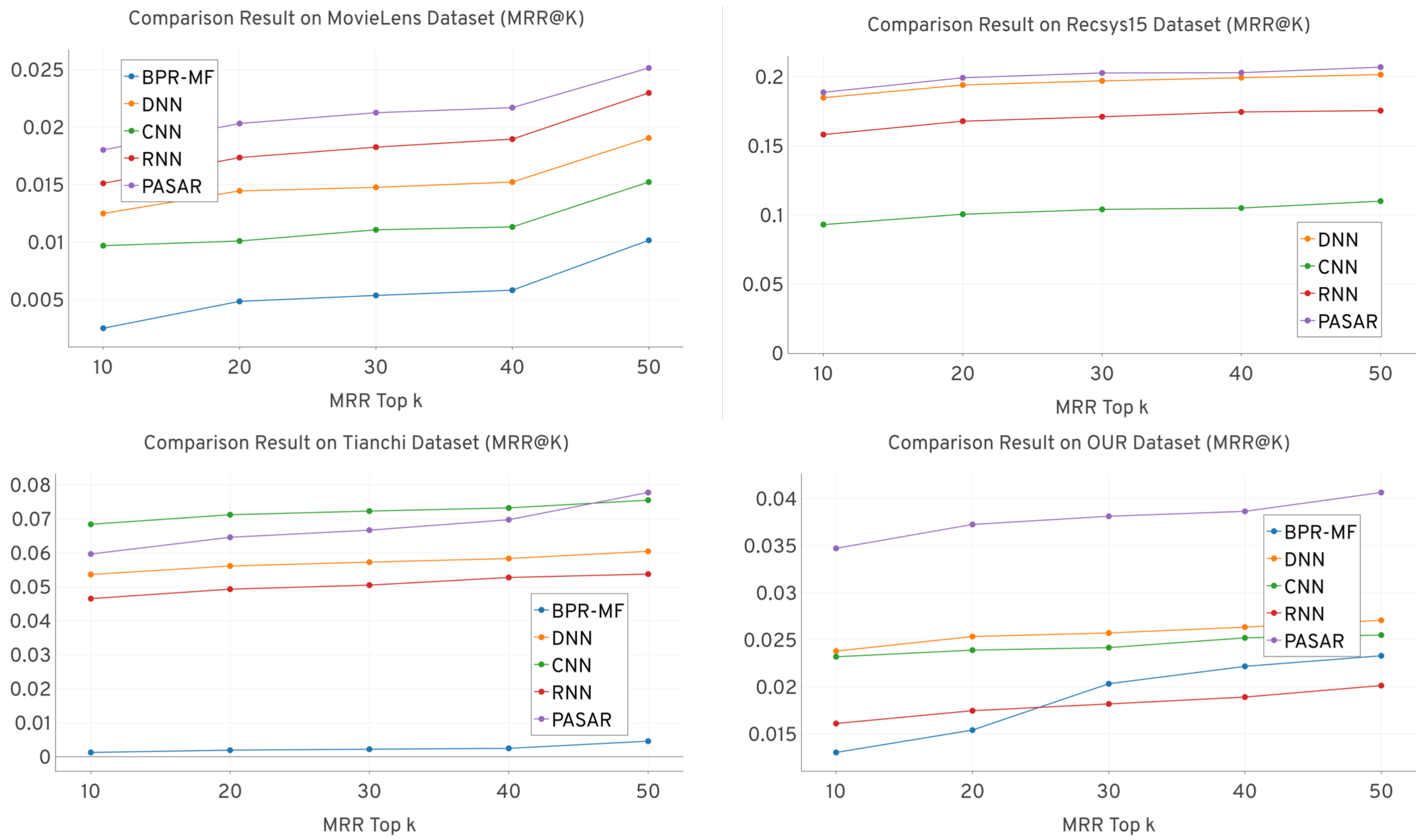
Evaluation: Comparison Results

Models	MovieLens		Recsys15		Tianchi		OURS	
	MRR@20	RECALL@20	MRR@20	RECALL@20	MRR@20	RECALL@20	MRR@20	RECALL@20
BPR-MF CF	0.004844	0.074627	/	/	0.001933	0.016234	0.015416	0.080431
YouTube DNN	0.014457	0.085271	**0.194101**	0.499136	0.056148	0.139335	**0.025355**	**0.103061**
WaveNet CNN	0.010098	0.054264	0.100597	0.33733	**0.071221**	**0.160209**	0.023910	0.100067
GRU4REC RNN	**0.017358**	**0.108527**	0.167908	**0.570426**	0.049316	0.127657	0.017474	0.058896
PASAR_user_att	0.012371	0.054264	/	/	0.041214	0.124636	0.015937	0.041002
PASAR_user_cat	**0.018451**	**0.100775**	/	/	**0.053976**	**0.138174**	**0.030365**	0.101227
PASAR_time_att	0.015988	0.038760	**0.199309**	**0.623005**	**0.057941**	0.146510	**0.021176**	0.067485
PASAR_time_cat	0.017539	0.038760	0.181273	0.589828	0.054056	0.140076	0.019368	0.061282
PASAR_time_user	**0.020321**	**0.100775**	/	/	**0.064585**	**0.204744**	**0.037259**	**0.106135**

Experimental Comparison Results -- shown are the MRR top 20 and Recall top 20 scores of four baseline models and five PASAR variants on four datasets.

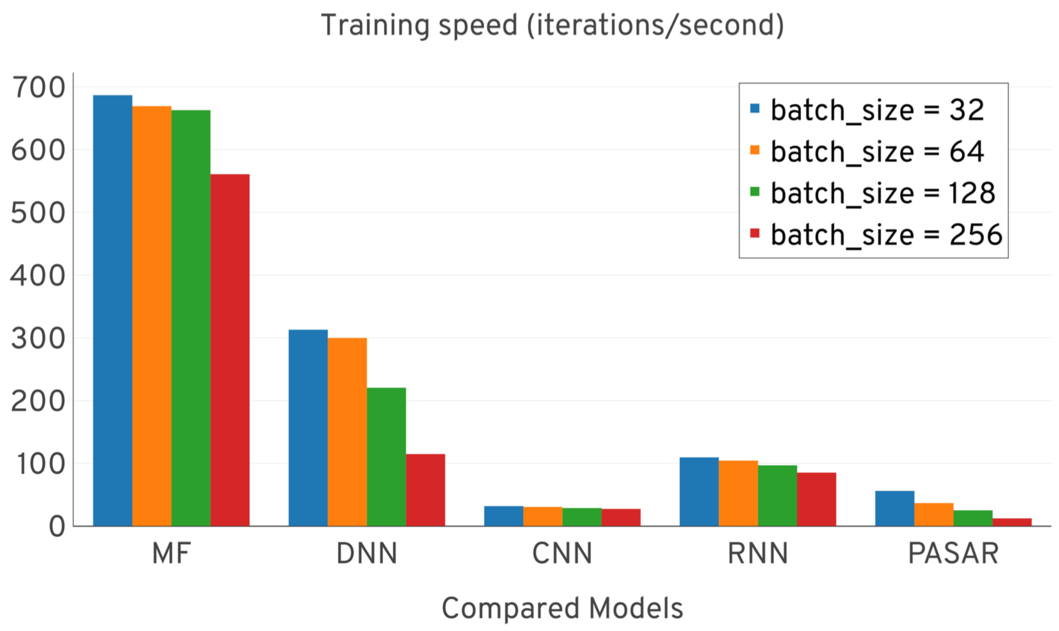
We highlight some focal improvements in bold and underline the best results.

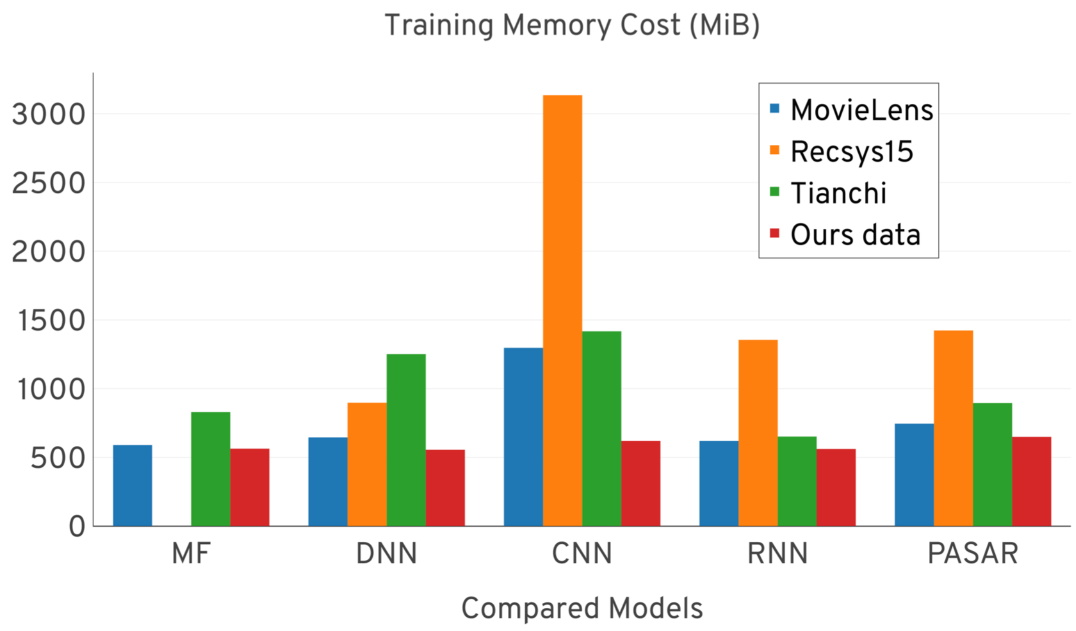
Evaluation: Comparison Results



The detailed MRR@10, MRR@20, MRR30, MRR40 and MRR@all results for each datasets

Evaluation: Comparison Results





Train speed time (iterations/second) and Training memory cost (MiB).

We did experiments on NVIDIA Tesla P40 GPUs. (172.20.190.45)

MF method is fastest and CNN method takes the most memory.

Our model is half slower than baseline RNN model and takes similar memory cost.

Conclusion:

- In this paper, we quantify, qualify and exploit the long–term user profile and short–term temporal dynamics for session–based RNN recommender systems.
- In particular, we propose a complete session–aware recommender system model, called "PASAR", to integrate intra–session and inter–session profiles for both users and items.
- We offer an extendable attention scheme to leverage temporal dynamics scheme exploiting more intra–session information so as to enhance session–based RS in time dimension.
- We also include long–term user profiles for session–based RS to learn the cross–session pattern and user favorite evolution in a seamless way.
- We demonstrate the improvement by our model design on four real–world datasets.

Discussion:

1. We can further optimize the attention scheme by local method, since the nearer item list attributes more for sequence predicting.
2. We can use grouped user representation to reduce dimensionality and accelerate the training process.
3. We can test on different support mechanism, like sparse and dense data, and make the model more robust.
4. More importantly, we can improve the negative sampling method and use better input embedding to solve the cold-start and low-rank problem.
5. CNN model is valuable to be further developed with signal processing techs.
6. We can use list-wise ranking paradigm for such top-k ranking models.

Summary:

1 Paper

2 Patents

Codes

4 Datasets

Many thanks to my mentors Weizhi and Chris.

Love JD.

Thanks

JD.COM