CUBC: A Generalized Representation Learning Method for User Behavioral Sequence

Yongqing Wang*, Haopeng Zhang*, Hao Gu†, Lingling Yi†, Huawei Shen* and Xueqi Cheng*

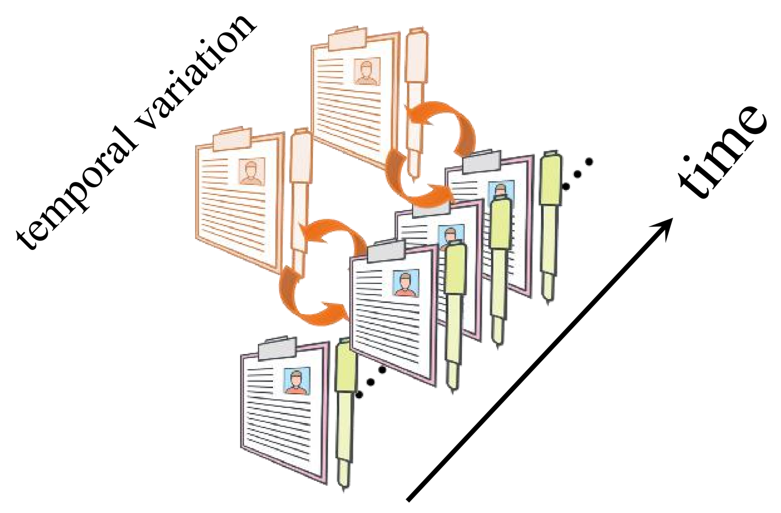
*Institute of Computing Technology, CAS, China †Weixin Group, Tencent Inc., China

Overview

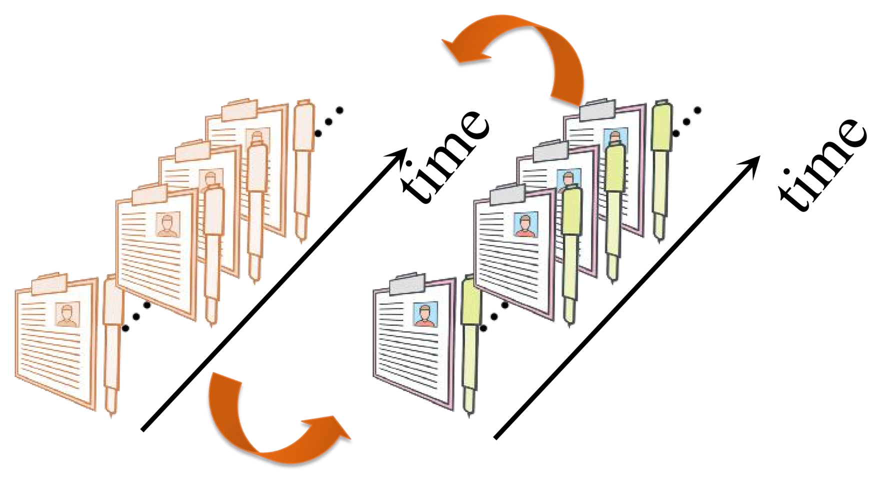
The objective of user behavior coding is to learn high-level representation over behavioral sequence which can be fed into downstream tasks, e.g., profiling and recommendation. The big challenge is that *can we learn generalized representation of behavioral sequence so as to support multiple downstream learning tasks*?

Our solution, CUBC, learns a generalized representation of user behaviors prefers to capture both **context- and content-level information**.

- **Context-level information**: Context-level information is sensitive to temporal variation in one sequence, which can be used to predict future or missing behaviors.

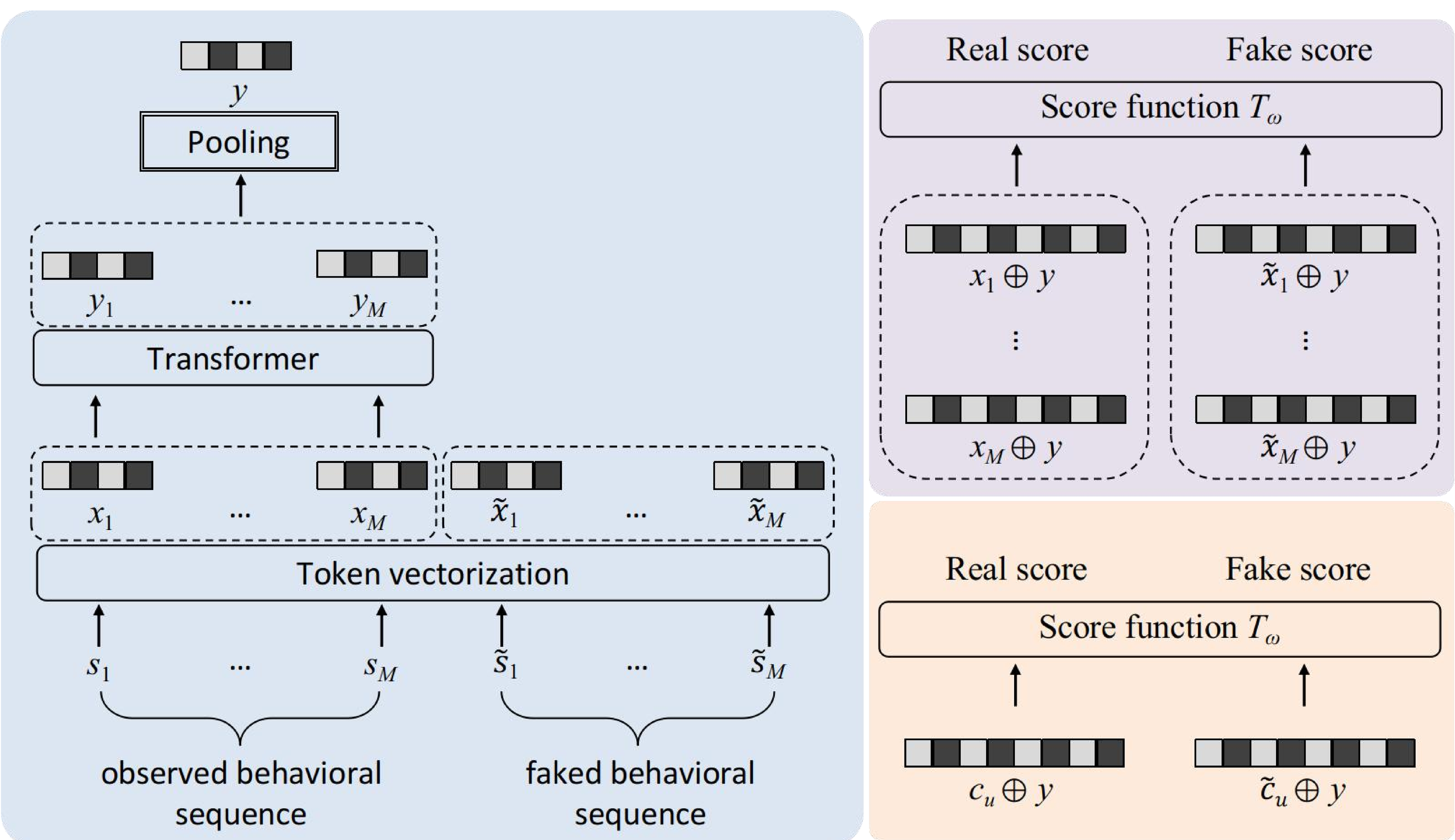


- **Content-level information**: Content-level information presents the sequence as whole, which can be generally applied to deduce stable characteristics relative to users.



Model

Based on mutual information (MI) maximization, Our basic learning framework is presented as belows, including representation learning, and context- and content-level mutual information.



Context-level MI maiximization:

$$\theta = \arg\max_{\theta} I(X; f_\theta(X))$$

θ : paramters
f : mapping function
X : input

Context-level MI maiximization:

$$\theta = \arg\max_{\theta} I(f_\theta(X); C)$$

C : side information (e.g., interests, age)

Experiments

Datasets Anonymized offline data from reading tracks of users on Wechat public subscription. Extracting 687,192 users and its corresponding behavioral sequences on reading activities during June 1 to June 30, 2019. For sequence modeling, the token in sequences refers to account ID read by users.

Baselines (1) Generative model: The learned representation is generated by objective on sequence completion. Generally, the last output is chosen for sequence representation; (2) End-to-end model: The model directly learn the mapping from input to downstream tasks; (3) Feature-based model: The downstream models are directly constructed by side information.

Results

Prediction performance on gender/age/next token prediction

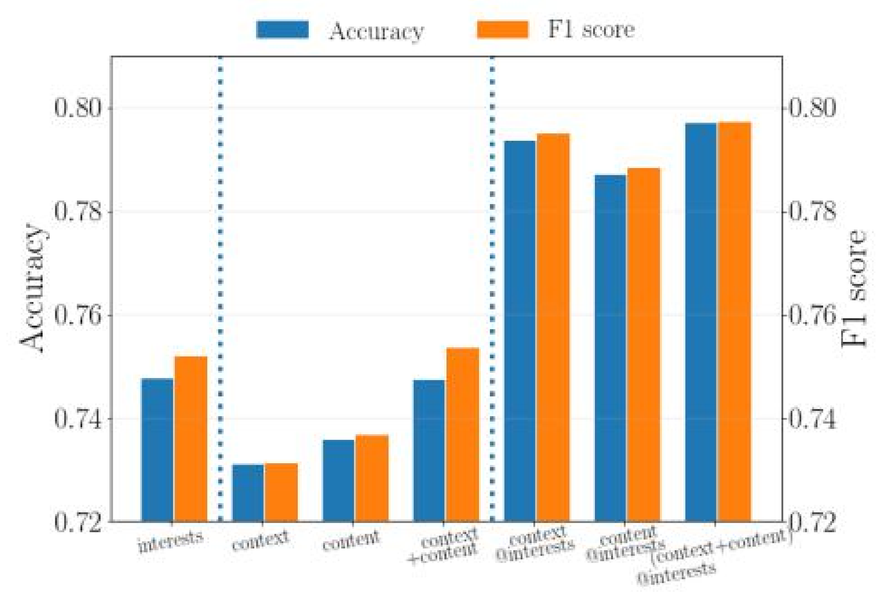
Model	Gender prediction				Age prediction				Next token prediction	
	Acc.	Precision	Recall	F1 score	Acc.	Precision	Recall	F1 score	nDCG@10	MRR
Generative (y_M)	0.6402	0.6236	0.7074	0.6629	0.6591	0.6912	0.5695	0.6245	0.6971	0.6805
End-to-end	0.7435	**0.7475**	0.7356	0.7415	0.7232	**0.7836**	0.6131	0.6879	0.9356	0.9067
Feature-based	**0.7465**	0.7381	0.7644	0.7510	0.6903	0.7072	0.6444	0.6743	0.5917	0.4783
CUBC (context)	0.7300	0.7296	0.7310	0.7303	0.7334	0.7407	0.7143	**0.7273**	**0.9692**	**0.9454**
CUBC (content)*	0.7349	0.7336	0.7379	0.7357	0.7226	0.7458	0.6715	0.7067	–	–
CUBC (context+content)	0.7463	0.7348	0.7710	**0.7525**	**0.7378**	0.7710	0.6730	0.7187	0.9672	0.9429

In all three tasks, the learned representations from CUBC outperform other comparative methods, even better than end-to-end model.

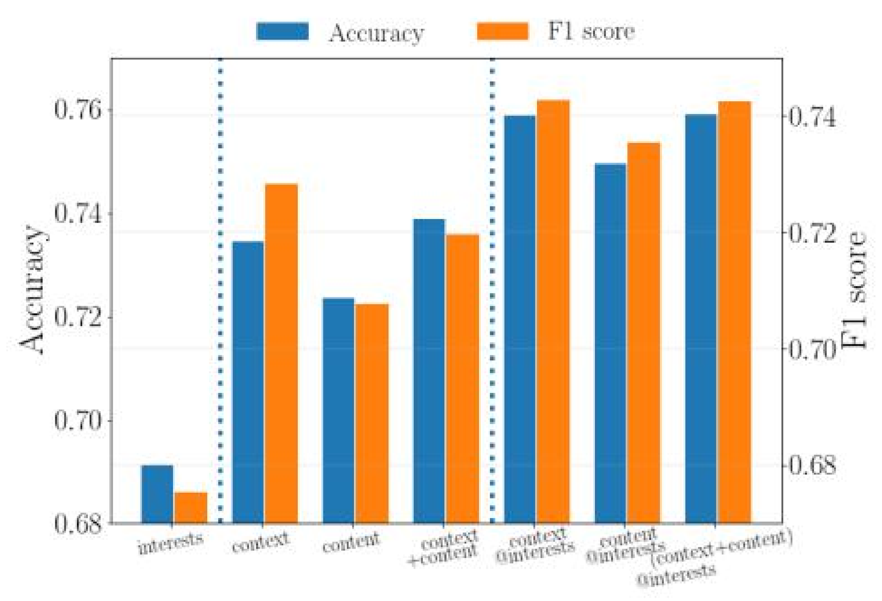
The Effect of side information

Gender prediction	Acc.	Precision	Recall	F1 score
Labels of gender	**0.7440**	0.7275	**0.7802**	**0.7529**
Labels of age	0.5636	0.5485	0.7204	0.6228
Interests of users	0.7349	**0.7336**	0.7379	0.7357
Sequence embedding	0.6752	0.6611	0.7190	0.6888

Age prediction	Acc.	Precision	Recall	F1 score
Labels of gender	0.5710	0.5910	0.4485	0.5100
Labels of age	**0.7233**	0.7294	**0.7060**	**0.7175**
Interests of users	0.7226	**0.7458**	0.6715	0.7067
Sequence embedding	0.7122	0.7333	0.6630	0.6963



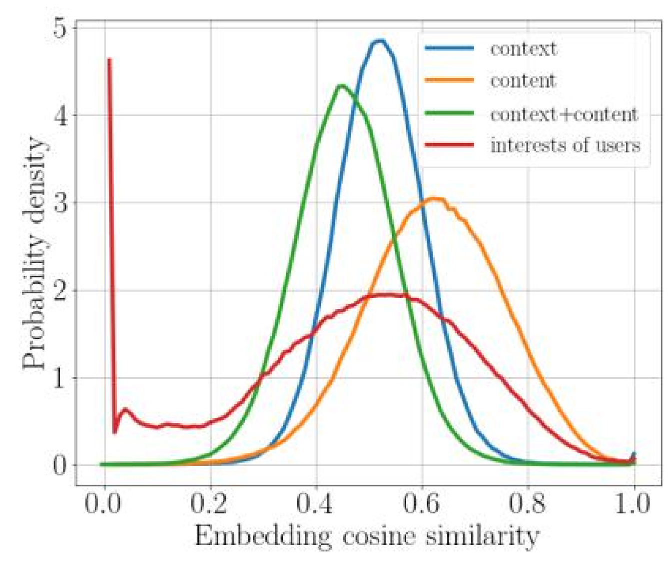
(a) Gender prediction



(b) Age prediction

High quality side information can optimize the learned representation. However, it is not equal to directly transform side information to learned representation.

Embedding similarity



Online A/B test

We also take A/B test experiments on online game recommendation production. After concatenating the learned representation from CUBC, the registration rate is lifted by +0.11%, +0.51%, 0.82% and +1.11% in four games.

Summary

- We propose an efficient coding method to learn sequence representation from users behaviors.
- Comprehensive experimental results prove the effectiveness of our proposed model on both offline and online scenarios.

Corresponding email: wangyongqing@ict.ac.cn
Homepage (first author): http://yongqwang.com