Learning to Predict the Departure Dynamics of Wikidata Editors (ISWC'21)

Guangyuan Piao (Maynooth University, Ireland**)**

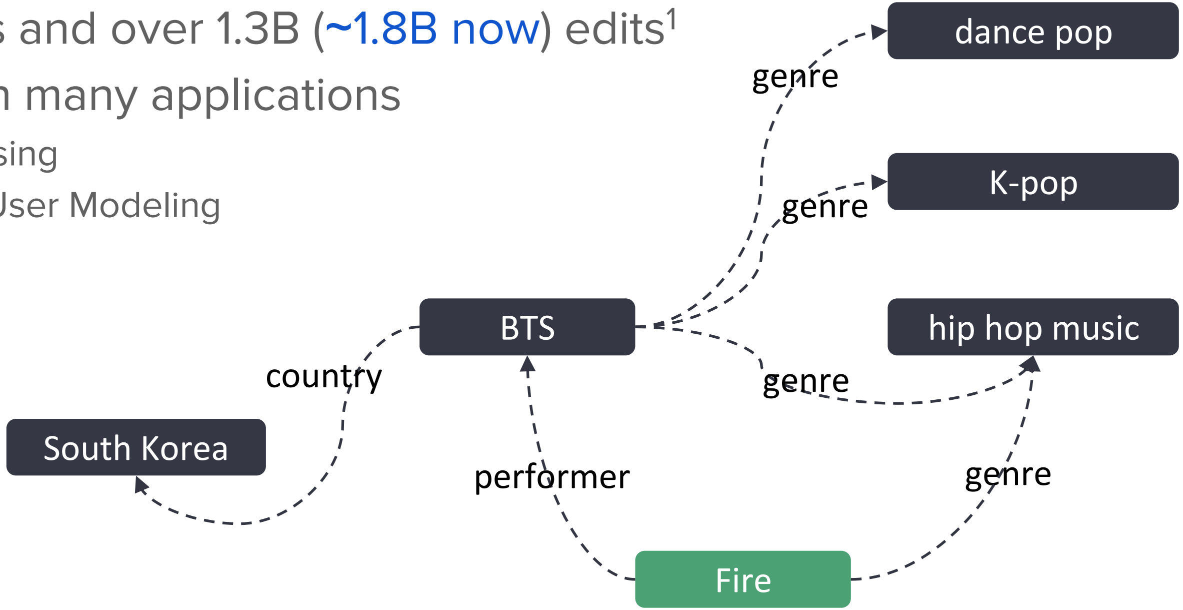
18/01/2023, Wikimedia Research/Showcase

Content

- **Motivation**
- Findings
- Wikidata Dataset
- Proposed Approach
- Results
- Conclusions

Motivation

- One of the largest open, free, multilingual knowledge bases
- 90M (~100M now) items and over 1.3B (~1.8B now) edits[1]
- Play an important role in many applications
  - Natural Language Processing
  - Recommender Systems, User Modeling
  - Life Sciences
  - ...



1. https://www.wikidata.org/wiki/Wikidata:Statistics

Motivation

- Editors (users) on Wikidata platform are critical to its success
- Understanding editing dynamics such as whether an editor will leave the platform is important but little attention has been given in the context of Wikidata, which is our focus of this study.

Motivation

- Editors (users) on Wikidata platform are critical to its success
- Understanding editing dynamics such as whether an editor will leave the platform is important but little attention has been given in the context of Wikidata, which is our focus of this study.

Problem formulation: $f(\mathbf{x}_u) \rightarrow y_u$

- $\mathbf{x}_u$ denotes a set of features based on the edit history of a user u,
- y_u is the class label indicating activeness of u (1 for inactive, 0 for active)

Content

- Motivation
- **Findings**
- Wikidata Dataset
- Proposed Approach
- Results
- Conclusions

Findings

Statistical features and **pattern-based** complement each other and ⬆ perf.

- total # of edits in the last 1 month
- distinct # of edited entities in the last 3 months
- …

Example: **rncv** for a pair of entities (i_1, i_2)

- i_2 is **r**e-edit of a previous entity
- i_2 is **n**ormal entity
- i_2 is a **c**ontinuous edit of i_1
- **v**ery fast edit from i_1

DeepFM model performs best with those features

Content

- Motivation
- Findings
- **Wikidata Dataset**
- Proposed Approach
- Results
- Conclusions

Wikidata dataset: **Dump of 2020-12-01**

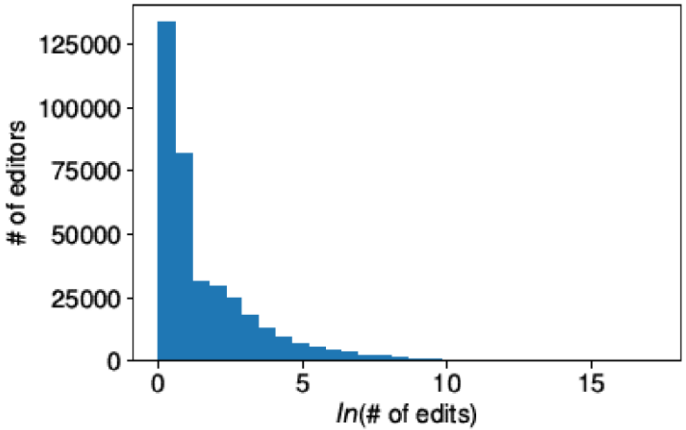
- Excluded edits from
  - anonymous users
  - bot accounts and administrators of Wikidata based on the open bot and admin lists

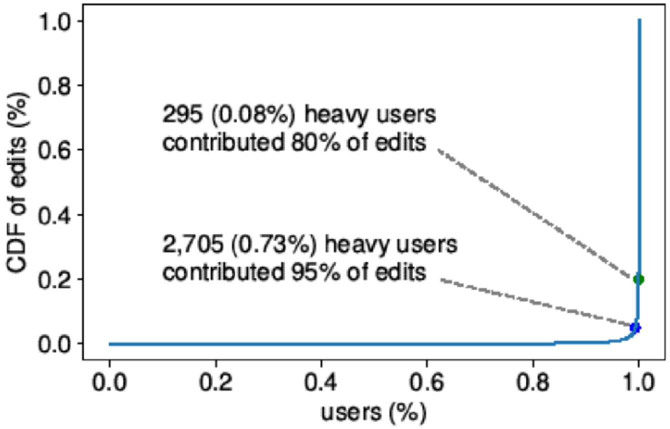
- **371,068** users & **519,121,793** edits

Username	Time	Entity ID	Edit action type

Wikidata dataset: **Observations**

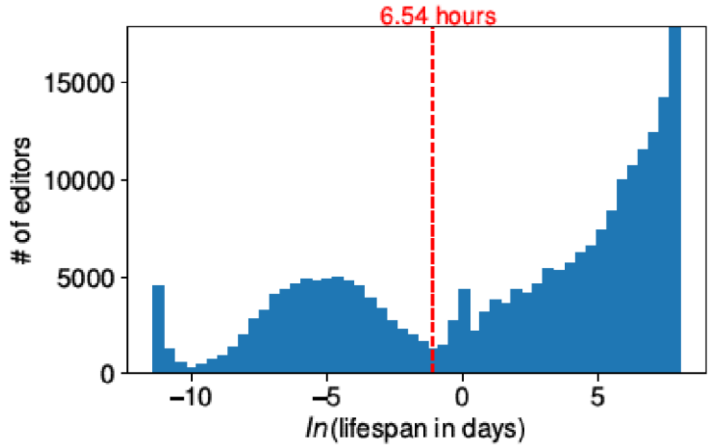
- a lot of users making a low number of edits
- a small number of heavy users making a high number of edits



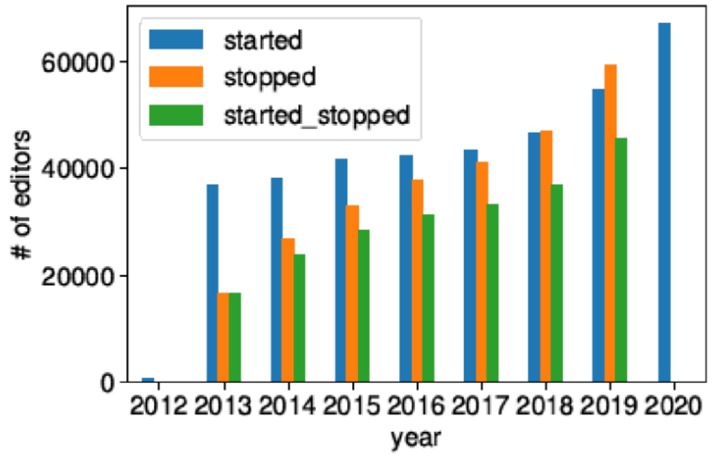


Wikidata dataset: **Observations**

- Lifespan: last edit – first edit (in days)
- 6.54 hours compared to 8 hours in Wikipedia[1]



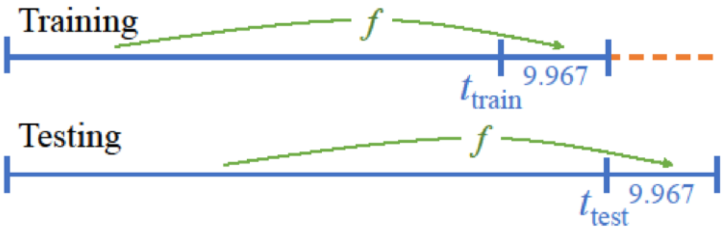
- The No. of newcomers is ⬆
- The No. of users who stopped is also ⬆
- Again, it is important to predict leaving editors and having additional efforts to keep those users



1. Zhang, D., Prior, K., Levene, M., Mao, R., van Liere, D.: Leave or stay: The departure dynamics of wikipedia editors. In: International Conference on Advanced Data Mining and Applications.

Wikidata dataset: **For our experiments**

- **Inactive user**: has not been editing any entity for 9.967 months (299 days)[1]



- For both training and testing, we limit users who are active before t_{train} and t_{test} to predict whether those active users will remain active or become inactive

Training (# of users)			Test (# of users)		
Active	Inactive	Total	Active	Inactive	Total
29,509	31,283	60,792	32,068	33,500	65,568

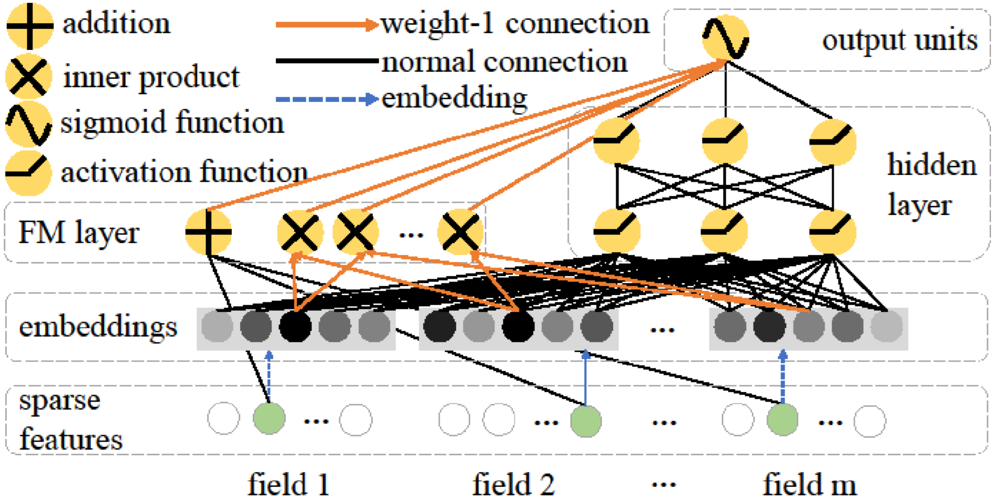
1. Sarasua, C., et al., The evolution of power and standard wikidata editors: comparing editing behavior over time to predict lifespan and volume of edits. CSCW'19.

Content

- Motivation
- Findings
- Wikidata Dataset
- **Proposed Approach**
- Results
- Conclusions

Proposed approach: **Overview**

- Statistical and/or pattern-based features
- A DeepFM classification model where those features are used as an input



Proposed approach: **Statistical** features

Statistical features (in *p* months)
• total # of edits
• distinct # of edited entities
• # of days between first and last edits
• diversity of edit actions
• diversity of entities
• diversity of day of week

$$p \in P = \{\frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 12, 36, 108\}$$

- # of days between first edit and prediction time
- # of days between the last edit and prediction time

Proposed approach: **Pattern-based** features

- From chronological sequence of each consecutive pair (i_1, i_2) of edited entities

- e.g., **rncv** for a pair of entities (i_1, i_2)
  - i_2 is a **r**e-edit of edited entities previously
  - and is a **n**ormal entity
  - **c**ontinuous edit of i_1 ,
  - **v**ery fast edit -- the time diff. between the two consecutive edits is less than 3 minutes.

Pattern-based features
• **r/n**: i_2 is re-edit or not
• **m/n**: i_2 is normal entity (starts with Q) or not
• If i_2 is a re-edit, **c/n**: i_2 is continuous edit or not • If not a re-edit, **z/o/u**: common classes of i_1 and i_2
• **v/f/s**: time diff. between two edits (e.g., very fast: v)

- Top 13 patterns for both classes (*active* and *inactive*), and *active* class
- of length $l\ \in \{1, 2, 3\}$, in total 13 * 2 * 3 = 78 features

SPMF

An Open-Source Data Mining Library

Content

Results: **Compared to other methods**

Method	*AUROC*	*F1*
GBT-Zh	0.8890	0.8205
kNN-Zh	0.8731	0.7935
RF-Sa	0.7647	0.7769
LR-Sa	0.7656	0.7795
SVM-Ar	0.8396	0.8029
DeepFM-Pattern	0.8786±0.0002	0.7992±0.0028
DeepFM-Stat	0.8928±0.0001	0.8247±0.0006
DeepFM-Stat+Pattern	**0.9561±0.0005**	**0.8843±0.0012**

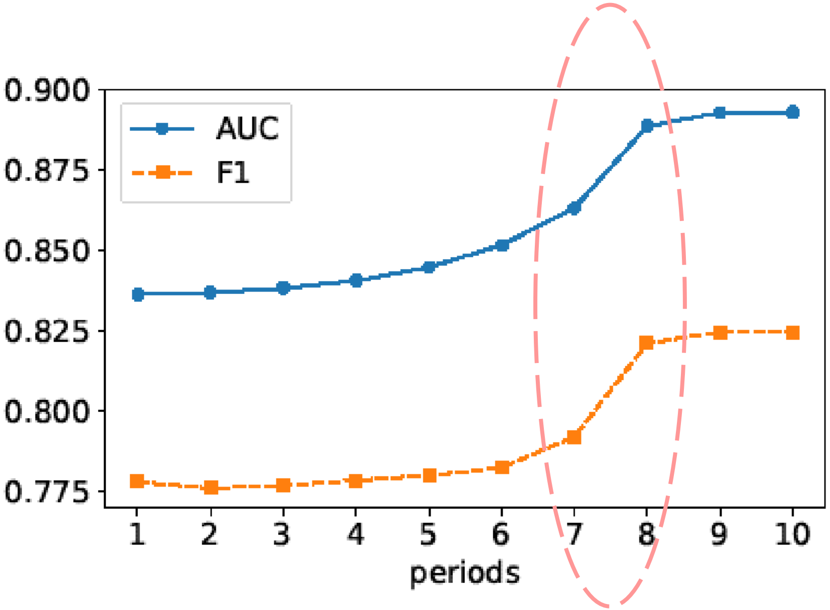
- DeepFM-Stat+Pattern provides the best performance in terms of *AUROC* and *F1*
- The two types of features - statistical and pattern-based ones - can complement each other and achieves the best classification performance

Results: **Efficiency of statistical features**

	AUROC (Improvement)	*F1* (Improvement)
GBT-Zh	0.8890	0.8205
GBT-Stat	**0.8918** (+0.32%)	**0.8225** (+0.24%)
kNN-Zh	0.8731	0.7935
kNN-Stat	**0.8898** (+1.91%)	**0.8172** (+2.99%)
SVM-Ar	0.8396	0.8029
SVM-Stat	**0.8640** (+2.91%)	**0.8040** (+0.14%)
DeepFM-Zh	0.8922±0.0001	0.8223±0.0029
DeepFM-Stat	**0.8928±0.0001** (+0.07%)	**0.8247±0.0006** (+0.29%)

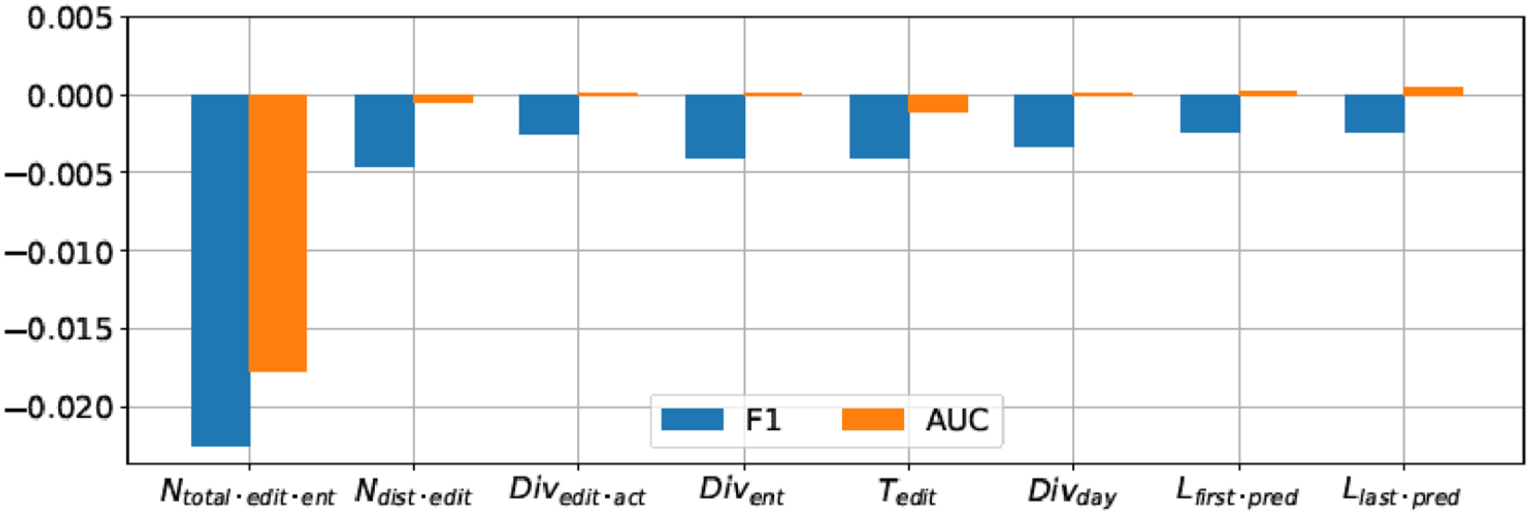
- Using our statistical features consistently achieves better *AUROC* and *F1* scores with those methods, which indicates the effectiveness of those features

Results: **Importance of periods**

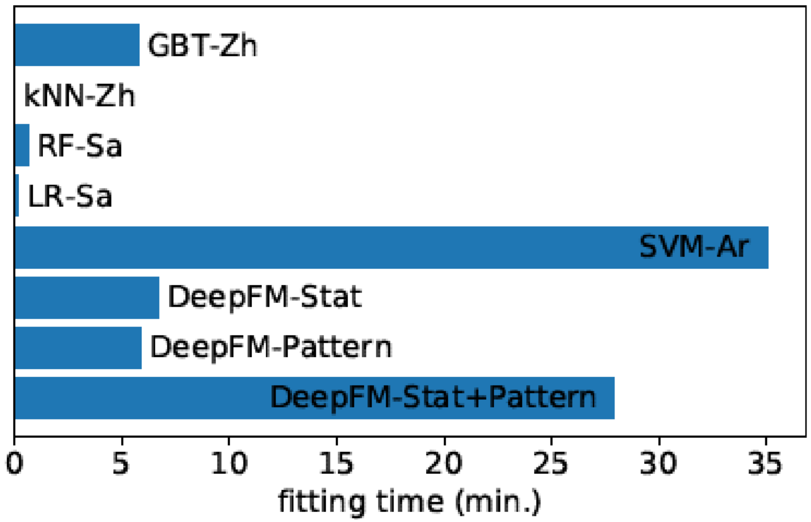


$$p \in P = \{\frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 12, 36, 108\}$$

Results: **Importance of features**



Results: **Time for training**



Content

- Motivation
- Findings
- Wikidata Dataset
- Proposed Approach
- Results
- **Conclusions**

Conclusions & Future work

- Investigated a set of proposed statistical and pattern-based features with DeepFM and built several adapted approaches from previous studies in the context of Wikipedia or different tasks
- DeepFM-Stat achieves the best *AUROC* and *F1* performance compared to other adapted methods when using either set of features
- Using both types of features can improve the performance significantly

- The promising results using DeepFM indicate other alternatives can be explored for the problem in the future
- Detailed analysis on pattern-based features can be explored