Fully-connected Attention Meta-path based Recommendation Model

Chunhui Wang
Department of Automation
Shanghai Jiao Tong University
chunhuiwang6@163.com

Ming Cheng
Department of Automation
Shanghai Jiao Tong University
chengming@sjtu.edu.cn

Abstract

Recommender system is playing a very important role nowadays, including music recommendation, advertisement recommendation, etc. A satisfactory recommendation will not only assist users in finding the items they want to find, but also will allow the online website to have more users so as to embrace more profits. There exist many algorithms to do the movie recommendation task, including Heterogeneous Information Network (HIN) and knowledge graph. Based on HIN, this paper proposed a fairly new model named FAMRM: Fully-connected Attention Meta-path based Recommendation Model, which extracts enough information of the relationship between users and movies (meta-paths) from the dataset and build a strong interaction between them through the full-connected structure. In addition, FAMRM considers the order of the recommended result by building a full-attention structure. After testing on the dataset MovieLens 100k, FAMRM can get a higher value of NDCG than recent existing works.

Keywords: *FAMRM, movie recommendation, full-connected structure, attention structure, meta-path*

1. Introduction

The recommendation system plays a very important role today. An appropriate recommendation system model can not only facilitate user's life (e.g. if an online shopping platform can reasonably recommend products that the user probably likes, it will greatly reduce the time for users to search for products), but also will provide sellers with huge profits (e.g. some advertising recommendations). Also, there are many types of applications for recommendation systems in various fields, including music recommendation, video recommendation, advertisement recommendation, etc.

Heterogeneous information network (HIN) is an important part of the recommendation system area. HIN can reflect some implicit and indirect relationship between the user and the item. The path that connects the user with the item is called the meta-path.

However, there are several problems with the proposed method of using meta-paths and HIN. When meta-paths are embedded, the feature spaces of users and items are actually inconsistent, so it is inappropriate to use their relative positions in their respective embedding spaces directly to describe the relationship.In addition after a variety of meta-paths having been weighted, actually the existed model does not take into account the user and the item to its indirect influence and also ignores the influence between the user and the item, so we choose to take a random walk way instead of using the vector distance in embedding space. Secondly, we use a full-connected connection layer in order to make full use of the user, the item, the path's influence on each other.

2. Background

There are many types of applications for recommendation systems in various fields. For example, [5] implements a recommendation system for YouTube videos, and [10] introduces the application in the field of music. Of course, in the wide application of recommendation systems, movie recommendation plays an unignorable role. Recommending movies appropriately not only allows the website to have more users, but also enables users with similar interests to have more connections with each other.

At the same time, with the development of computer science, many algorithms related to recommendation systems have emerged. [7] and [8] make recommendations based on collaborative filtering and decision tree, respectively. Collaborative filtering collects user's preference data and calculates the similarity of each item, including Euclidean Distance [21], Pearson Correlation Coefficient [1] and so on. However, it has to involve the cold start problem [2], which will make the recommendation system's effect for new users not good. Decision tree [18] is a classic machine learning algorithm that can predict output values based on user-related data. However, when it comes to chronological issues, decision tree algorithm does not work well. In recent years, with the development of deep learning [14], knowledge graph [16] and other technologies, the usage of them in the field of recommendation systems has become more and more widespread. As for deep learning, [4] proposed a deep candidate generation model and describe a separate deep ranking model for YouTube video recommendation. Also, RankClus, a new model based on Heterogeneous Information Network (HIN), is proposed in [22]. Meanwhile, the knowledge graph, as a way of representing the knowledge encoded in data as well as a tool to reason on them in order to extract new and implicit information, has also been implemented in recommendation system in the field of music [15]. In 2018, [11] proposed MCRec, a neural co-attention model based on meta-path, and tested it on MovieLens 100k as well as other datasets, with a good result having been achieved.

Based on HIN [23] as well as taking the shortcomings of the existing network [11] into consideration, this paper proposed a fairly new model - FAMRM: Fully-connected Attention Meta-path based Recommendation Model. FAMRM fully extracted the information of the dataset and established a variety kinds of meta-paths. Besides, FAMRM considered both direct and indirect connection between the user and the item by establishing a "fully-connected" structure. What's more, FAMRM considered the impact of all the elements in users, items and meta-paths and then proposed a "full-attention" structure. After testing the model on several datasets, FAMRM can get very satisfactory results.

3. THE PROPOSED MODEL

In this section, we will introduce the new proposed model called Fully-connected Attention Meta-path based Recommendation Model(FAMRM).

3.1. Model overview

Figure 3 is the framework of our proposed FAMRM model. At the beginning of model, we model various meta-paths to get different vectors. These vectors come from different ways of connection between the user and the item, such as UMUM,UMTM,UMTMUM and UUUM. Next we use an attention layer to combine these vectors into one path vector.The initial user and the movie's one-hot vector pass through the lookup layer to become the embedding vector. At this point,we get three vectors, which are the weighted path vector, the embedding user vector and the embedding movie vector. So we're going to use these vectors to learn the attention layer for each of these vectors. Then we multiply the three vectors by their corresponding attention layers.The new three vectors are then combined into one vector and passed through an MLP layer to get the final result.

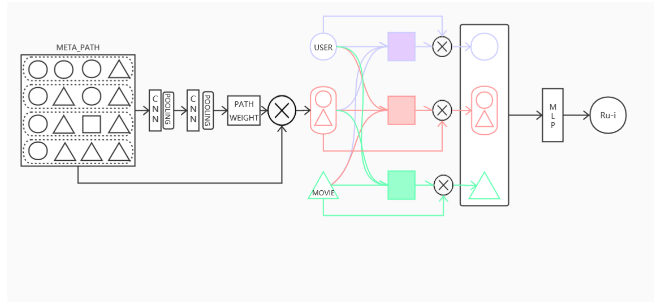

Figure 1. The overall framework of FAMRM.

3.2. Meta-path

This is one of the important parts of our algorithm innovation.The data preprocessing method metapath2vec adopted in the previous McRec method which firstly obtains embedded vectors of different categories. And then in the random walk, it picks the nodes that are closest to the current node. However, there is a big problem with this method. The embedding space of features of different categories is not consistent, so it is impossible to measure the similarity of the two nodes by the relative distance of positions in the two spaces. So we use the traditional random walk to solve this problem. By using different modes, we obtain a number of vectors used to characterize different meta-paths. In order to obtain the deep features, we used CNN and maxpooling layers to obtain the deeper features.

$$p_{c1...cn} = CNN(p_{o1...on})$$

$$p_{1...n} = maxpooling(p_{o1...on})$$

At this point, we get multiple vectors that describe the node similarity. We use an attention layer to obtain vector weights of different paths.

$$\alpha_{p_{1\to n}} = f_0 \begin{bmatrix} p_1 & ... & p_n \end{bmatrix}$$

We choose the two highest correlation paths, as inputs to meta-path.

$$\alpha_{p_{1\to 2}} = f_1 \begin{bmatrix} p_1^{'} & p_2^{'} \end{bmatrix}$$

The weights are then redistributed again to obtain a meta-path vector that is weighted by the attention layer.cAnd then we multiply the weight of attention and two paths are got.

$$P = \alpha_{p_{1\to 2}}^{T} * \begin{bmatrix} p_1 & p_1 \end{bmatrix}$$

3.3. Fully-connected Attention

3.3.1 User And Item Embedding

For a single user and a single item, both of these vectors are both one-hot vector. We use a lookup layer to transform it into a embedding vector. Then we get the embedded matrix Q, P through machine learning algorithm, which is in the formula ,$u\epsilon\mathbb{R}^{|u|*1}$,$i\epsilon\mathbb{R}^{|i|*1}$,$Q\epsilon\mathbb{R}^{|u|*d}$,$P\epsilon\mathbb{R}^{|u|*d}$

$$Xu = Q^T * u$$

$$Yi = P^T * i$$

3.3.2 Fully-connected Attention layer

This is the core of the model we proposed. We can now get the embedding user and item vectors and the processed path vector. So what we started out with was equal weight fusion of these three vectors.

$$Cu \rightarrow i = Xu + Yi + P$$

But it doesn't show the different contirbution of these three vectors, so we want to add an attention layer to solve this. For users and item vectors, we consider the direct impact of the path on them as well as the indirect impact of the embedding item and user vector on them respectively, which means we consider the user, path, and item in the attention layer at the same time. In the formula ,,$\alpha i, \alpha i, \alpha i \epsilon \mathbb{R}^{|d|*1}$

$$\alpha u = f_2(\begin{bmatrix} Xu & Yi & P \end{bmatrix})$$

$$\alpha i = f_3(\begin{bmatrix} Xu & Yi & P \end{bmatrix})$$

And as for the meta-path, which in this case is an unprocessed path vector, we take the impact of the user, the item into consideration. We input the user, the path, and the item vectors into the attention layer.

$$\alpha p = f_4(\begin{bmatrix} Xu & Yi & P \end{bmatrix})$$

The main purpose of this is to consider that not only the synthetic path vector may affect the attention layer of user and item, but also the item and user vector may affect the attention layer of paths due to their indirect relationship. Those are important to the path attention layer, whihc means not taking this effect into account in the previous model is a loss of information.

3.4. MLP layer

So now we have three weighted vectors. To map it to a lower dimensional space, we use the MLP method to reduce it to one dimension. Finally, the value used to characterize the similarity between the item and the user is obtained.

$$X_{u \to i} = \begin{bmatrix} {\alpha_u}^T * Xu & {\alpha_i}^T * Yi & {\alpha_p}^T * P \end{bmatrix}$$

$$\gamma_{u \to i} = MLP\left(X_{u \to i}\right)$$

In this way, for the original two one-hot vectors (users and items), we finally get the similarity of the two of them. The higher the value, the more likely the user likes the item.

4. EXPERIMENTS

In this section, we evaluate the performance of FAMRM.

4.1. Dataset and Experimental Setup

4.1.1 About the dataset

The dataset we test our model on is MovieLens. MovieLens describes the relationship between users and movies and the detailed expression of it is shown in the table below (all numbers indicate the number of items (users and movies)).

Relations(A-B)	A	B	A-B
User-Movie	943	1682	100000
User-User	943	943	47150
Movie-Movie	1682	1682	82798
Movie-Genre	1682	18	2861

In the data, we select four kinds of meta-paths in total: UMUM, UMTM, UUUM and UMTMUM. As mentioned in section 3.2, the four kinds of meta-paths we select can extract enough information about the relationship between users and movies from the dataset. We will also compare the performance of different algorithms (ItemKNN [20], BRP [19], MF [13], NeuMF [9], $SVDFeature_{hete}$ [3], $SVDFeature_{mp}$ [3], HeteRS [17], FMG_{rank} [24] and MCRec [11]) on this dataset.

4.1.2 Experimental Setup and Implementation Details

We implement the FAMRM using the python library of Keras. For our model, we randomly initialize model parameters with Gaussian distribution and optimize the model with Adam [12]. Specifically, the dimension of user and item embeddings are set to 128. Besides, all the experiments are conducted on a machine with GPU: NVIDIA GeForce RTX 2080Ti.

4.2. Experimental Results

The comparison results of our proposed model FAMRM and other models are shown in the table below.

(1) As can be seen from the table, FAMRM is almost better than all the other models on the dataset MovieLens. The result indicates the effectiveness of FAMRM on the task of top-N recommendation.

(2) From the table, we can see that FAMRM can get the prediction accuracy as **0.3327**, the value of recall as **0.2183** and the value of NDCG as **0.7208**. Although the accuracy of prediction lowers a little, the value of NDCG improves a lot, indicating that we can predict correctly for the first few categories.

(3) Also, from the table we can conclude that most of HIN based methods (SVD, FMG) behave better than CF methods (ItemKNN, BPR and MF), indicating the usefulness of heterogeneous information. Besides, we can see that NeuMF behaves best among all the baselines and the possible reason is that NeuMF utilizes the multi-layer perceptron

Model	Prec@10	Recall@10	NDCG@10
ItemKNN	0.2578	0.1536	0.5692
BRP	0.3010	0.1946	0.6459
MF	0.3247	0.2053	0.6511
NeuMF	0.3293	0.2090	0.6587
$SVDFea_{hete}$	0.3171	0.2021	0.6445
$SVDFea_{mp}$	0.3109	0.1929	0.6536
HeteRS	0.2485	0.1674	0.5967
FMG_{rank}	0.3256	0.2165	0.6682
MCRec	0.3394	0.2227	0.6866
$FAMRM_{pri}$	0.3421	0.2245	0.6900
FAMRM	**0.3327**	**0.2183**	**0.7208**

[1] $FAMRM_{pri}$ is a variant of FAMRM which employs our meta-path innovation.

[2] FAMRM is our complete model

to model the complex interaction between users and items. In addition, this result indicates the excellence of deep neural network in building complex interactions in the field of recommendation [11].

(4) Lastly, FMG_{rank} performs best among all HIN based methods, indicating that effective features from similarity evidence have been learned. In addition, we can see that SVDFeature does not work very well. After analysis, we think it is possibly because the learned embeddings from metapath2vec [6] don't provide very useful information.

The possible reasons that we can get higher value of NDCG are as follows:

- The kinds of meta-paths we bulid and choose extract enough features from the original data.
- The full-attention structure we bulid considers the impact of all the elements in users, items and meta-paths.
- The full-connected structure we bulid considers both direct and indirect interaction between users and movies.

Because NDCG is considered as the most important evaluation index in the field of recommendation, the higher value of NDCG indicates the suitability and correctness of our model.

5. Detailed Analysis

5.1. The Effect of Different Meta-paths

Because the way we deal with meta-paths is that we do a random walk and we pick the two that have the highest impact, so we did an experiment. This experiment compared the effects of two different combinations and we found that the ones that had the highest proportion of attention had the best effect. Therefore, it is important to analyze the importance of different combinations of meta-paths. The definition of four kinds of meta-paths are as following:

- UMUM (user-movie-user-movie): If a user loves a movie and this movie is also loved by another user, then what the second user loves will also be loved by the first user, possibly.
- UMTM (user-movie-type-movie): If a user loves a movie and this movie belongs to one type, then the other movies also belonging to this type will be loved by this user, possibly.
- UUUM (user-user-user-movie): If a user loves a movie and there exist several other users who are "similar" to this user (the similarity is defined by KNN algorithm), then what these users love will also be loved by this user.
- UMMM (user-movie-movie-movie): If a user loves a movie and there exist several other movies which are similar to this movie (the similarity is defined by KNN algorithm), then all these movies will also be loved by this user, possibly.

In order to obtain the knowledge of the effect of different kinds of meta-paths, we choose two kinds of meta-paths from all of the four (UMUM, UMTM, UUUM, UMMM) each time, and train the network. The result is shown in the figure below:

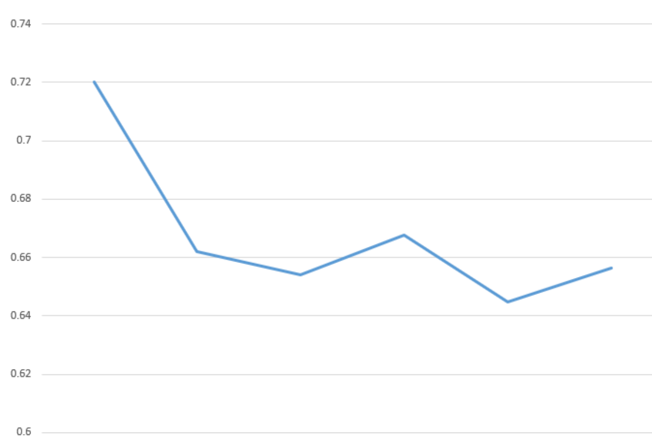

Figure 2. The effect of different kinds of meta-paths combination.

(1) From the figure, we can see that the choice of "UMUM, UMTM" behaves best among all the 6 choices, with the value of NDCG as 0.72. After analysis, we think the reason is that the meta-paths got by KNN algorithm (UUUM and UMMM) will effect the final result of the model. Because KNN is to define the similarity between two items in this case, and the result given by KNN is not the actual

similarity between these two, which means there exists an error of the usage of KNN algorithm to build meta-paths. Also, we can see that the result of the combination "UUUM, UMMM" get the worst result (0.6449), which is consistent with our analysis above: because both these two kinds of meta-paths are built using KNN algorithm, the extraction of information from the dataset is not accurate.

(2) Also, we can see from the figure that if the combination involves UUUM, the result will not be very satisfactory, mostly: "UUUM, UMUM" can get the result of 0.6678, but "UMTM, UMUM" can get the result of 0.72; "UUUM, UMMM" can get the result of 0.6449, but "UMUM, UMMM" can get the result of 0.6563. After our analysis, we think that it is because the similarity between users is not easy to get, even we use KNN algorithm. Because there exist lots of factors which will enable two users to be "not similar", (such as occupation, age, etc.), it is full of difficulties to measure the similarity between users, causing the worse result of the model.

5.2. Full-attention Structure

As mentioned above, the full-attention structure we build considers the impact of all the elements in users, movies and meta-paths, leading to the higher value of NDCG. Because we allocate different weights inside the vector of users, movies and meta-paths when doing the full-attention work, we can enable the order of final result to be more reasonable if we do attention process on all of the three vectors instead of on just part of them.

5.3. Full-connected Structure

The full-connected structure includes the direct interaction between users and movies as well as indirect interaction between them, which means we build the connection through meta-paths. After we consider the direct interaction between users and movies, it is easier to find the relationship between them, which will improve the final result, obviously. However, the indirect interaction between them cannot be ignored because through meta-paths, we can also extract useful information, which is also very helpful to build the relationship between users and movies. Besides, we use several dropout processes to weaken the interaction between users and movies in order to prevent the phenomenon of overfitting. All in all, through the full-connected structure as well as the dropout process, we can not only build stronger relationship between users and movies, but also can avoid the overfitting phenomenon.

5.4. Other Discussions

Although we get a very satisfactory value of NDCG, the accuracy of prediction doesn't improve. We think that it is possibly because the full-attention structure and the full-connected structure will enable the interaction between users and movies to be too strong, causing the overfitting phenomenon to appear in a way. However, because NDCG is thought to be the most important index in the field of recommendation, we think it is more important to recommend to user "what they like mostly" instead of just recommend "what they like" to them. Recommending "what they like mostly" to users will enable them to get the result which is most similar to what they expect, which is also consistent with the aim of recommendation system.

6. Conclusion

Our paper proposes a new approach called the Fully-connected Attention meta-path based Recommendation Model (FAMRM).This method improves the previous sampling method and obtains a new metapath sampling method. In addition, considering the influence of the other two parts on every part and the difference of the final result of each part, we adopted the method of combining the three vectors to learn the attention layer of each vector.

Various experimental results prove the superiority of our model in recommendation quality and validity. We believe that the FAMRM is a useful method to use in the field of recommendation system.

There are still some problems in the present work, which we hope can be solved in the future research. The first one is that currently the extraction of meta-paths still relies on manual processing of selected datasets. In the future, we hoped that the graph model can be used to automatically segment the graph and generate embedding vectors obtained by random walk from the small segmented graph. The second one is that the last result is generated by MLP, but this learning method is still linear. Maybe other learning methods can be used to reduce the dimension more accurately.

References

[1] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.

[2] JesúS Bobadilla, Fernando Ortega, Antonio Hernando, and JesúS Bernal. A collaborative filtering approach to mitigate the new user cold start problem. *Knowledge-Based Systems*, 26:225–238, 2012.

[3] Tianqi Chen, Weinan Zhang, Qiuxia Lu, Kailong Chen, Zhao Zheng, and Yong Yu. Svdfeature: a toolkit for feature-based collaborative filtering. *Journal of Machine Learning Research*, 13(Dec):3619–3622, 2012.

[4] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198. ACM, 2016.

[5] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He,

Mike Lambert, Blake Livingston, et al. The youtube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 293–296. ACM, 2010.

[6] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 135–144. ACM, 2017.

[7] Michael D Ekstrand, John T Riedl, Joseph A Konstan, et al. Collaborative filtering recommender systems. *Foundations and Trends® in Human–Computer Interaction*, 4(2):81–173, 2011.

[8] Nadav Golbandi, Yehuda Koren, and Ronny Lempel. Adaptive bootstrapping of recommender systems using decision trees. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 595–604. ACM, 2011.

[9] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and TS Chua. Neural collaborative filtering in: Proceedings of the 26th international conference on world wide web, 173–182. In *International World Wide Web Conferences Steering Committee*, 2017.

[10] Wendell Hicken, Frode Holm, James Clune, and Marc Campbell. Music recommendation system and method, Feb. 17 2005. US Patent App. 10/917,865.

[11] Binbin Hu, Chuan Shi, Wayne Xin Zhao, and Philip S Yu. Leveraging meta-path based context for top-n recommendation with a neural co-attention model. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1531–1540. ACM, 2018.

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[13] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, (8):30–37, 2009.

[14] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[15] Sergio Oramas, Vito Claudio Ostuni, Tommaso Di Noia, Xavier Serra, and Eugenio Di Sciascio. Sound and music recommendation with knowledge graphs. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(2):21, 2017.

[16] Enrico Palumbo, Giuseppe Rizzo, and Raphaël Troncy. Entity2rec: Learning user-item relatedness from knowledge graphs for top-n item recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, pages 32–36. ACM, 2017.

[17] Tuan-Anh Nguyen Pham, Xutao Li, Gao Cong, and Zhenjie Zhang. A general recommendation model for heterogeneous networks. *IEEE Transactions on Knowledge and Data Engineering*, 28(12):3140–3153, 2016.

[18] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[19] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 452–461. AUAI Press, 2009.

[20] Badrul Munir Sarwar, George Karypis, Joseph A Konstan, John Riedl, et al. Item-based collaborative filtering recommendation algorithms. *Www*, 1:285–295, 2001.

[21] Huifeng Sun, Yong Peng, Junliang Chen, Chuanchang Liu, and Yuzhuo Sun. A new similarity measure based on adjusted euclidean distance for memory-based collaborative filtering. *JSW*, 6(6):993–1000, 2011.

[22] Yizhou Sun, Jiawei Han, Peixiang Zhao, Zhijun Yin, Hong Cheng, and Tianyi Wu. Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, pages 565–576. ACM, 2009.

[23] Xiao Yu, Xiang Ren, Yizhou Sun, Bradley Sturt, Urvashi Khandelwal, Quanquan Gu, Brandon Norick, and Jiawei Han. Recommendation in heterogeneous information networks with implicit user feedback. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 347–350. ACM, 2013.

[24] Huan Zhao, Quanming Yao, Jianda Li, Yangqiu Song, and Dik Lun Lee. Meta-graph based recommendation fusion over heterogeneous information networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 635–644. ACM, 2017.