The Analysis and Design of the Job Recommendation Model Based on GBRT and Time Factor

Pengyang Wang, Yingtong Dou and Yang Xin

Beijing University of Posts and Telecommunication

INTRODUNCTION

In this paper, we first give a comprehensive summary of models and algorithms applied in three online job recommender systems and point out the advantages and disadvantages of these models. Then we introduce a job recommendation model based on Gradient Boosting Regression Tree and time factors (T-GBRT). The T-GBRT model aggregates the time factors into the GBRT to predict personal preferences and adds time factor weight to topK rankings, with a neighbor based filtering trick in reducing the amount of calculation. At the end of the paper, the model performs the best in the experiment with four criterions, comparing to other three models, which proves the efficiency of the new model.

For the designing of recommender systems for online recruitment, there are two main models having been proposed.

Collaborative Filtering (CF) [1]. There are user based CF and item based CF[2] and being applied in the scenario where the number of users is bigger or the number of items is bigger respectively. R. Rather, K. Bradley and B. Smyth[3] had proposed two approaches in CASPER system. One method is based on the user based CF; it recommends target users the jobs which their similar users like. CASPER gives recommendation according to the ranking. Calculating the similarity of users and jobs to evaluate the compatibility of users and jobs. W. Hong et al. clustered the users into three clusters according to their information and take three different recommendation policies[4]. One of the method is based on user based CF.

Content based recommendation[5]. Its main idea is to calculate the jobs which are similar to the jobs having been interacting with users according to the users' interacting record and then recommend them to the users. The measurement of its similarity is based on the content of the job itself which has many types i.e. the texts. CASPER[3], the iHR of W. Hong[4] et al., and the PROACTIVE of Lee[6]

DETAILS OF THE MODEL

1. STRUCTURE

Our model includes four main parts, the feature extraction, the chosen of alternative set, the training and prediction of GBRT (Gradient Boosted Regression Trees) model and the TopN recommendation generated by the weight of time factor. The model is shown in the Fig. 1

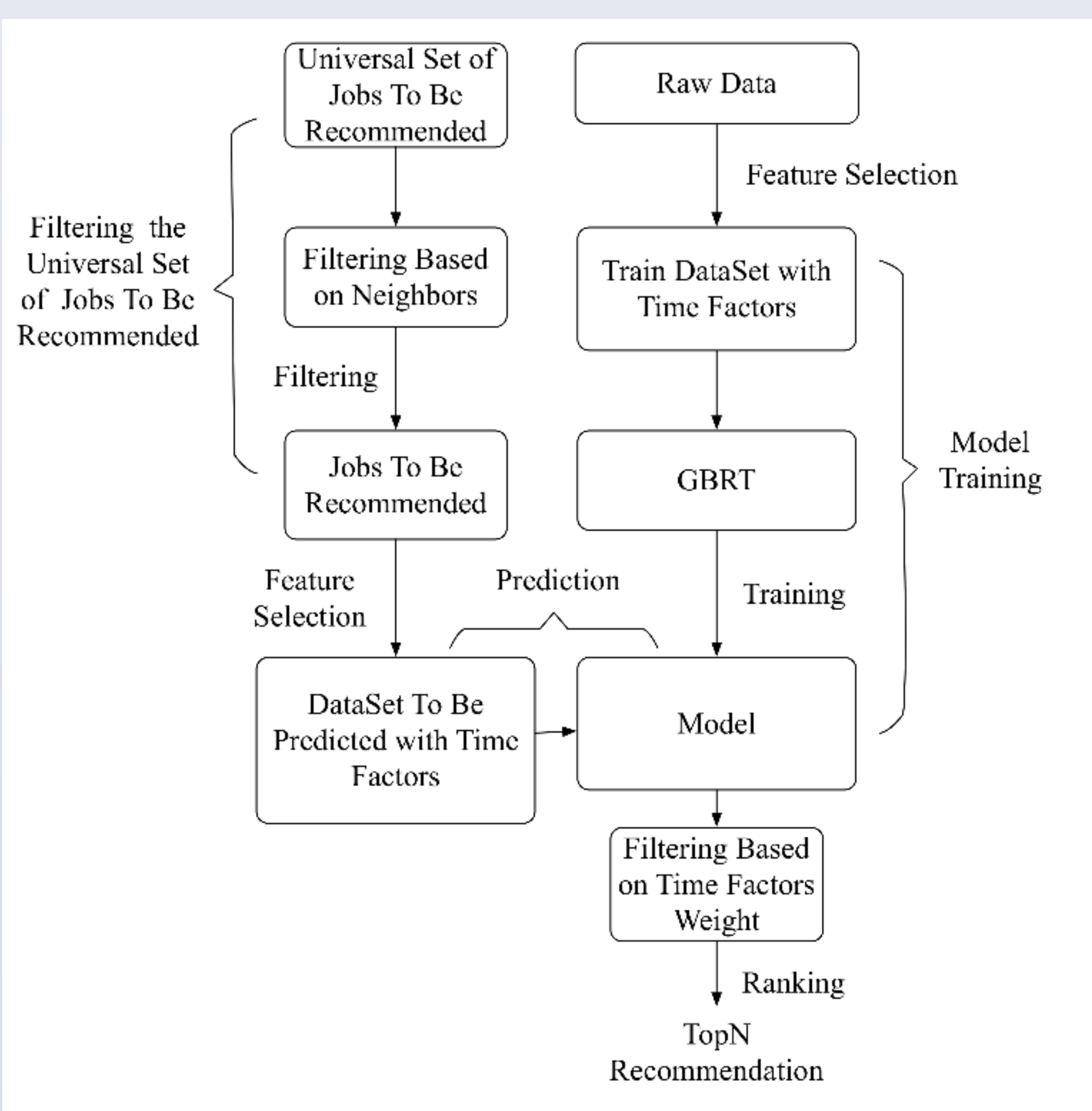

Fig. 1.This figure introduces details the structure of the T-GBRT model.

2. TIME FACTOR BASED FEATURE

The future interaction will be influenced by the past behavior online. Taking online shopping as example, if the user bought item A and liked A, the user will show his/her preference to B when the user buys item B which is similar to A after a month. It could be seemed as the influence from A to B. We need to consider the attenuation effect and aggregation effect when we describe the feature.

The attenuation effect means that the interaction which is longer to now has less influence on the present. If the influence is represented by function F, F is smaller while the time duration is larger.

The aggregation effect means the jobs having been interacted with the user all have influence on the further interaction. We need to accumulate the influence of all the interacted jobs when we quantify the aggregation effect.

The time factor based feature is defined as

$$feature_{time\_factor}(u,j) = \sum_{\Delta t} F(\Delta t)(\sum_{i} G(w_{i,j}, r_{u,j}))$$

u means the specific user, i means the jobs having interacted with user u. $w_{i,j}$ means the similarity between job i and job j, $r_{u,i}$ means the rating of user u to job i. Δt means the time duration from the time when user u has interaction with job i to now. $F(\Delta t)$ represents the time attenuation effect and it decreases as the Δt increases. $G(w_{i,j}, r_{u,i})$ represents the influence from once interaction between user u and job i to user u and job i without consideration of time factor.

The user u has no direct relation to job j but they are both related to the job i. The relation between user u and job i is the rating $r_{u,I}$ and $w_{i,j}$ connects job j with job i. The job i is the bridge between user u and job j. Their relationship can be shown as Fig. 2.

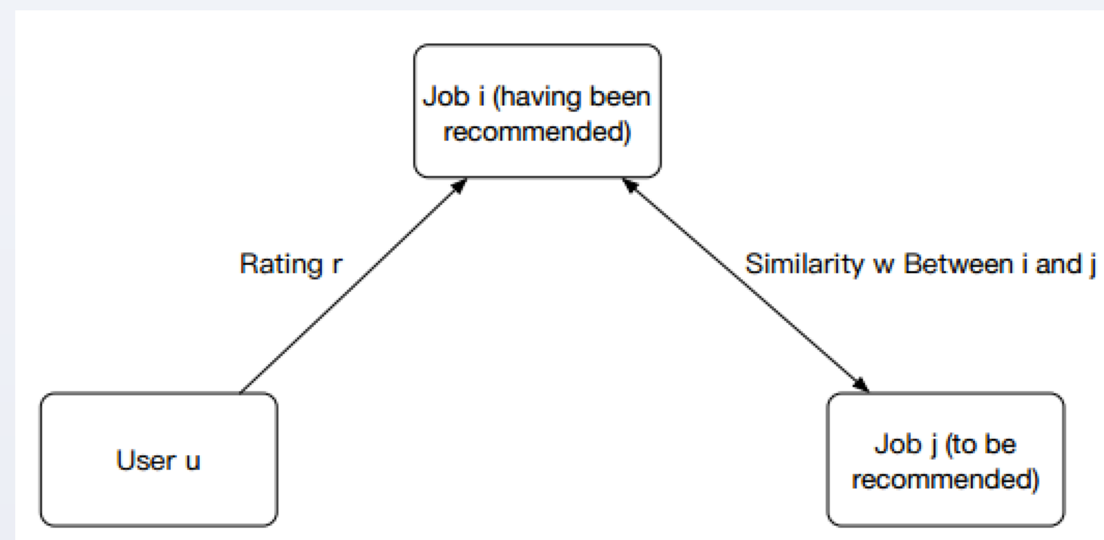

Fig. 2.This figure describe the relationship between user u, job i(having been recommended) and job j(to be commended).

EXPERIMENT

1. DATA

The data in the experiment are from the 2016 ACM Conference on Recommender Systems challenge. The data are provided from the sponsor www.xing.com which is a job hunting website. The data has four main parts, the recommendation generated by the website users themselves (impression.csv), the user information (users.csv), the job information (items.csv) and the user behavior log (interactions.csv). The time span is from the 34th week to the 46th week in 2015. We choose the data from the 34th to the 45th week. The training set is the data from the 34th to the 44th week. The data of the 45th week is the offline testing data. Here, in the training process, the number of total job hunters is 741650 and the number of jobs is 1500000; in the test process, the number of target job hunters is 133796 and the scale of available awaiting job lists is 370925.

2. EVALUATION METHOD

The testing method refers to the testing method of 2016 ACM Conference on Recommender Systems challenge. We calculate the marks based on the scoring function score(S, T)[7].

3. COMPARISION OF RESULTS

We could get that the T-GBRT is better than other three models in total score, precision, recall rate and hit numbers. For the precision, the improvement of TopN at the first half part and last half part is quite a lot and it shows that our model could optimize the user experience. UBCF and IBCF only consider the interaction between users and items but not the influence of content and time. CBR add the content but not the time factor. Our T-GBRT model integrate the interaction history, neighbor, time and performs better in each criterion.

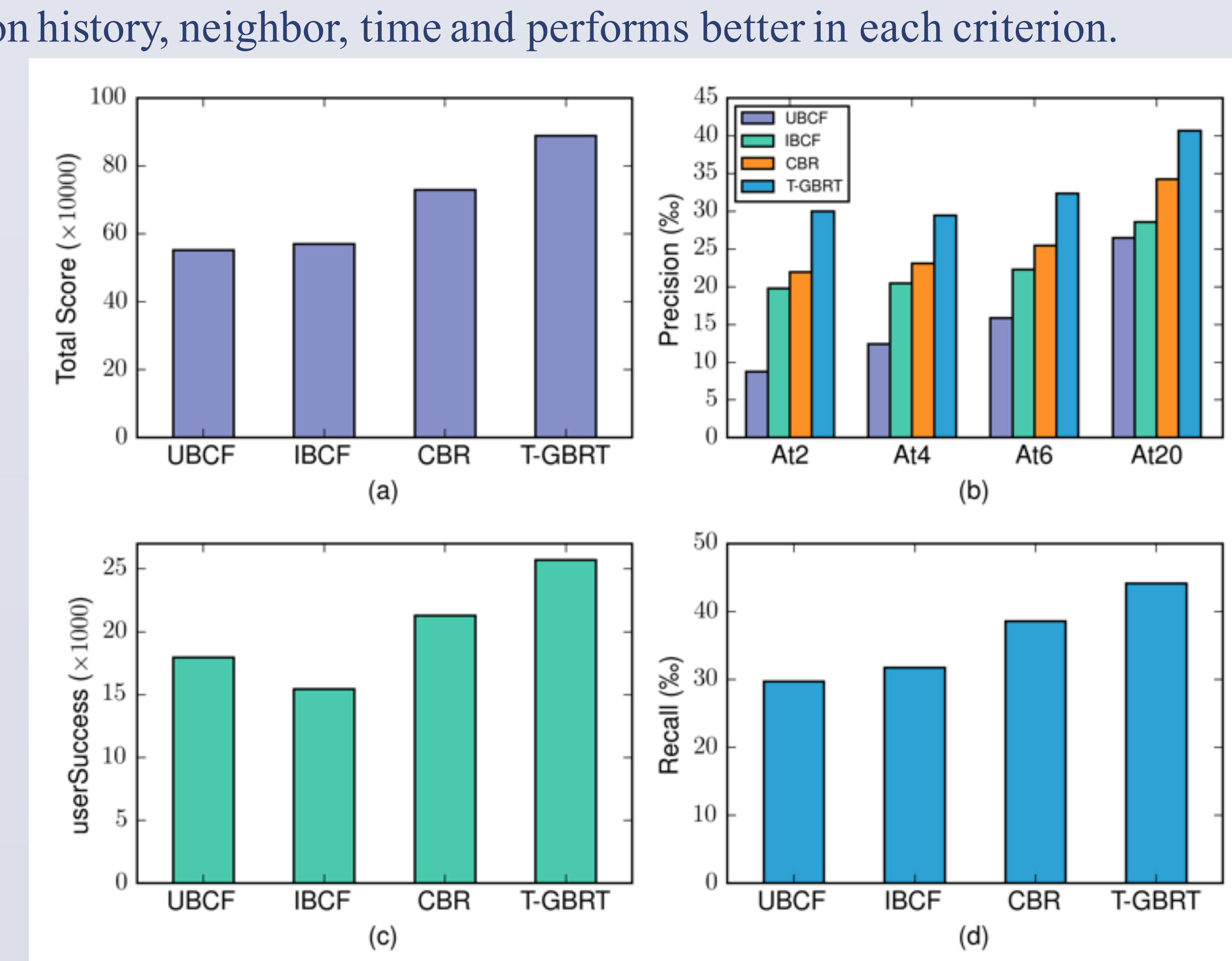

Fig. 3.(a) illustrates the total score of the four model respectively, according to the evaluation method above. (b) gives the precision of four models at the first 2, the first 4, the first 6 and the first 20, respectively. (c) gives the number of success recommendation of four models. (d) gives the recall of four models. (a)~(d) correspond to the numerical values in the TABLE IX. , respectively.

REFERENCE

[1] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens - An Open Architecture for Collaborative Filtering of Netnews.," *CSCW*, pp. 175–186, 1994.

[2] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms.," *WWW*, pp. 285–295, 2001.

[3] R. Rafter, K. Bradley, and B. Smyth, "Personalised retrieval for online recruitment services," *In Proceedings of the 22nd Annual Colloquium on Information Retrieval*, 2000.

[4] W. Hong, S. Zheng, H. Wang, and J. Shi, "A Job Recommender System Based on User Clustering.," *JCP*, vol. 8, no. 8, 2013.

[5] M. J. Pazzani and D. Billsus, "Content-Based Recommendation Systems.," *The Adaptive Web*, vol. 4321, no. 10, pp. 325–341, 2007.

[6] D. H. Lee and P. Brusilovsky, "Fighting Information Overflow with Personalized Comprehensive Information Access: A Proactive Job Recommender," presented at the Third International Conference on Autonomic and Autonomous Systems (ICAS'07), 2007, pp. 21–21.

[7] Https://github.com/recsyschallenge/2016/blob/master/EvaluationMeasure.md