Proceedings of the Tenth International AAAI Conference on
Web and Social Media (ICWSM 2016)

Who Are Like-Minded: Mining User Interest Similarity in Online Social Networks

Chunfeng Yang
The Chinese University of Hong Kong
Hong Kong, China
yc012@ie.cuhk.edu.hk

Yipeng Zhou
Shenzhen University
Shenzhen, China
ypzhou@szu.edu.cn

Dah Ming Chiu
The Chinese University of Hong Kong
Hong Kong, China
dmchiu@ie.cuhk.edu.hk

Abstract

In this paper, we mine and learn to predict how similar a pair of users' interests towards videos are, based on demographic, social and interest information of these users. We use the video access patterns of active users as ground truth. We adopt tag-based user profiling to establish this ground truth. We then show the effectiveness of the different features, and their combinations and derivatives, in predicting user interest similarity, based on different machine-learning methods for combining multiple features. We propose a hybrid tree-encoded linear model for combining the features, and show that it out-performs other linear and tree-based models. Our methods can be used to predict user interest similarity when the ground-truth is not available, e.g. for new users, or inactive users whose interests may have changed from old access data, and is useful for video recommendation.

1 Introduction

How well can an online social network (OSN) service use various user profile information, including demographic and social interaction information, to predict user interest similarity for recommendations for related services (such as video service), is the question we study in this paper. The conventional approach for video recommendation is collaborative filtering (CF) (Su and Khoshgoftaar 2009), based on analyzing past user access activities (or more explicit feedback). CF can tell us which users (or items) are more similar based on video access information. But such information is not available for inactive and new users. For this reason, user interest similarity prediction based on other (demographic and social) information, if accurate, is very helpful.

Without content categorization (via CF or other methods), computing and comparing two users' interest similarity directly from the items (videos) they accessed in the past can be misleading. For example, two users with similar interest may happen to access different videos of the same categories, generating no (or little) overlap to indicate their similarity, which is a wrong conclusion. In our study, we rely on a tag-based video categorization system already available in the data source[1]. We adopt two methods to use the tagging information to compute user interest similarity. One method is based on the frequency of common tags between users, called Popular Tag based Profiling (PTP); and the other method takes into account the representativeness of the tags, called Representative Tag based Profiling (RTP).

Given PTP and RTP, we can systematically derive interest similarity of active users (given we have their video access data) and use it to study how different demographic (e.g. age, gender, location) and social (friendship, community membership, and interaction information) features can predict the similarity based on PTP and RTP. We also studied the similarity of a user's interest with her past; and with the average of all users. The results show that we can effectively identify the usefulness of different features, and some of them can help infer user similarity effectively. After that, we apply different machine learning models, including linear models, tree-based models, and a hybrid model developed by ourselves, to study how to best use the available features to infer user interest similarity for users without past access data. The results show that our hybrid method out-performs the other learning models. Finally, to demonstrate the benefit of our results in real applications, we apply the similarity inferred from our models to video recommendation, and compare it to several benchmarks. The experimental results validate the strength of our prediction model and the tag-based user profiling scheme in real applications. While PTP-based similarity results in more accurate recommendations, RTP-based similarity can produce more diverse recommendations.

2 Data and Background

This study is based on data from our collaborator - Tencent Inc[2]. Tencent is a major social network provider in mainland China, running a platform for its instant messaging (QQ) service, many online games, a social network and social media (WeChat) service, online Video service and others.

In this paper, we use the following user profile data:

- Demographic information: we utilized three types of information among others- gender, age, and location.
- Social relationship: we considered friendship, social interaction and group membership. Given two users, friendship is mutual. Social interaction means instant messag-

[1]Each video can have multiple tags. The set of tags are manually generated by multiple editors. This is considered a folksonomy approach.

[2]The first author has interned at Tencent.

ing between friends. Group membership is the set of QQ groups[3] joined by a given user.

- Interest data: we used users' video access data over time.

The sets of users and videos are denoted by $\mathcal{U}$ and $\mathcal{I}$, respectively. Total number of users is $|\mathcal{U}|$. We use u and v as the indices of users, and m and n as the indices of videos. The gender, age and location of a user u are represented by g_u, a_u and l_u. User u's friend set is $\mathcal{F}_u$ and we denote the friendship between user u and v as $F(u,v)$, where $F(u,v)=1$ if $u \in \mathcal{F}_v$ and $v \in \mathcal{F}_u$. The QQ groups joined by user u is denoted as $\mathcal{G}_u$. Interaction (messaging) between friends on a certain day is represented by $m_d(u,v)$, where d is the index of the day (if we index the current day as 0, then the past day is -1, and the like). The viewed video set of user u is $\mathcal{I}_u$[4], which in fact is the traditional video-based profile of user u.

3 User Interest Profiling based on Tags

In traditional CF based on user-item access data (not rating data), the simpler memory-based approach to compute user similarity does not do well if data is sparse. Two users with similar interest may have accessed quite different items (videos). This can be remedied with model-based CF (Koren, Bell, and Volinsky 2009), but that can be quite compute-intensive, and requires recomputation as new users and content come in. Thus, we use a tag-based profiling system which is a quite common practice in industry.

As is typically done, Tencent Video employed many "editors" dedicated to viewing and labeling videos with one or more tags from a predefined tag vocabulary (in order to avoid colloquial and noisy tag usage). For each video, tags used by more than a certain proportion of editors are kept. There are on the order of 30 thousand tags in the tag vocabulary. Some examples are "visually intriguing", "reality show" and "mixed feelings". We then generate tag-based user profiles by fusing the user-video consumption data and video-tag relations. Thus, users consuming the same video will be labeled with the same set of tags.

Another important point is that we are interested in users' *current* interest similarity that is more accurately captured by only recent viewing behaviors (see more justification and discussion in Sec. 4). We define active users as ones who have viewed one or more videos in a recent time window (a day is used in our study); the other users are considered inactive. New users are by default inactive. It is quite common for users who are once active to become inactive.

We further define some tag-related notations. The set of tags is represented by $\mathcal{T}$. We use i and j as the indices of tags. The tag set of video m is denoted as $\mathcal{T}_m$. The tag set of user u is $\mathcal{T}_u = \{i | i \in \mathcal{T}_m, m \in \mathcal{I}_u\}$, and the set of users who have tag i is $\mathcal{U}_i = \{u | i \in \mathcal{T}_u\}$.

[3]Each social group contains typically 50 to 100 users, sharing some common interest.

[4]Notations related to users' viewing behaviors all refer to one day's data unless stated otherwise, and for simplicity, the time stamp is omitted if there is no ambiguity.

Definition 1 *(Popular Tag based Profiling, PTP). The weight of tag i is proportional to the number of videos labeled by tag i and viewed by user u during a certain period. The user profile, obtained by aggregating the tag sets of videos viewed by u, is denoted by*[5]

$$\mathbb{T}_u^P = \left\{ \left(i : w_i^P\right) | w_i^P = \left| \{m | i \in \mathcal{T}_m, m \in \mathcal{I}_u\} \right| \right\} \quad (1)$$

To identify more informative and representative tags, we propose another tag-based profiling method by penalizing tags that are very popular among user profiles.

Definition 2 *(Representative Tag based Profiling, RTP). Besides user u's individual preference of tag i, RTP also considers the occurrence of tag i in all users' tag lists. The user profile is represented as*

$$\mathbb{T}_u^R = \left\{ \left(i : w_i^R\right) | w_i^R = w_i^P * log_2 \frac{|\mathcal{U}|}{|\mathcal{U}_i|} \right\} \quad (2)$$

For each user profiling method, namely, PTP, RTP and video-based profiling, we use cosine similarity measure to calculate the similarity, denoted by $S^P(u,v)$, $S^R(u,v)$, and $S^I(u,v)$ respectively. See (Yang, Zhou, and Chiu 2016) for the detailed formulas.

4 Correlation Study and Inferring Interest Similarity Between Users

We conducted extensive empirical studies to explore correlations between various user features and interest similarity, and try to seek key features that influence interest similarity. The detailed empirical results are in (Yang, Zhou, and Chiu 2016). From the empirical studies, we obtained 10 discriminative features belonging to three categories, namely demography (i.e., gender pair, age pair, location pair), social relations (i.e., friendship, ratio of common friends, number of common groups, monthly message count, monthly messaging days) and interest (i.e., past long-term similarity, individuality).

In this section, we attempted to investigate two issues: 1) among different machine learning models, which one is best suited for our prediction tasks; 2) the effectiveness of different categories of features and their combinations. To illustrate them, we conduct two series of experiments: 1) with all the selected features, we fit them with various algorithms to learn different models; 2) for each feature combination, we construct a predictive model using pruned decision tree. For each series of experiments, we examine the performance of interest similarity prediction in both PTP and RTP cases for the classification (**two users are similar or not**) and regression tasks (**the similarity value between two users**).

From users who were active on the target day (August 30th, 2015), we randomly selected one million user pairs for each experiment. We used seventy percent of the samples for training and the rest as the testing data. Moreover, ten-fold cross validation was used in model training. For the classification task, we binarized the similarity values with a

[5]Since each tag in the profile has a weight, we use a dictionary-like structure to represent the profile, that is, $w_i^P = \mathbb{T}_u^P[i]$. The same structure is adopted for profiles in RTP.

certain threshold (the mean value of interest similarity) so that we could predict whether two users are similar or not. To evaluate the performance, we utilized the evaluation metrics of area under the ROC curve (**AUC**) which is insensitive to label imbalance. And for regression, to avoid the specific values of regression error, we used reduced ratio of mean absolute error (**reduced ratio of MAE**) as the performance metrics which is calculated by comparing the error with that achieved by a constant estimator using the mean value of similarity in the training set.

4.1 Experiments of Various Machine Learning Models

Considering the heterogeneity of these features, we propose a hybrid tree-encoded linear model. In this model, we firstly used gradient boosting decision tree (GBDT) (Ye et al. 2009) to encode the features by converting the original features into some binary features. Each encoded feature corresponds to a region jointly described by multiple ranges of original feature values. Furthermore, to complement the possibly missed linear relations between the output and the original features, we applied both the encoded features and original features into a regularized linear modelto reduce redundancy among those features. The details of the hybrid model can be found in (Yang, Zhou, and Chiu 2016). For comparison, we used two linear models and two tree models to fit the data. The linear models consist of simple linear models and l1-regularized linear models, and the tree models comprise pruned decision tree models and ensemble tree models. The complete performance results of different models in similarity prediction are in (Yang, Zhou, and Chiu 2016).

From the results, under both PTP and RTP, the hybrid tree-encoded linear model achieved the best performance in the classification and regression tasks, meaning that it best fit our problem. In fact, the tree-encoded features could achieve feature combinations automatically so as to capture the non-linear and multi-feature (i.e., membership combined with friendship) relations. Different from traditional ensemble tree models, namely random forest and GBDT which assigns a weight for each sub-tree, our hybrid tree-encoded linear model will learn a weight for each leaf node of the sub-trees and for each original feature.

We also applied the hybrid tree-encoded linear model to predict video-based similarity defined in Sec. 3 with our full feature set, which will be used in Sec. 5.

4.2 Experiments of Different Feature Combinations

To test the predictive ability of different features, we trained pruned decision tree with different feature combinations. See (Yang, Zhou, and Chiu 2016) for the complete results.

As shown in the results, each category of features would contribute to the interest similarity prediction on its own. And with more selected features, the prediction performance is better. Although various social relations are apparently correlated with interest similarity as show in the empirical studies, they are not available to many user pairs. Moreover, the results show that for user pairs with partially available information, such as, only social and demographic information, we could still adopt the predictive models to improve the interest similarity estimation.

5 Apply Our Findings to Recommendation

To demonstrate the practical value of the proposed prediction algorithms and the tag-based profiling scheme, we applied the predicted results to video recommendation which recommends a list of videos matching the user's video interests. For traditional recommendation algorithms, such as CF and content-based filtering, it is difficult to provide accurate recommendation for users with little or no recent interest information, which is known as the cold start problem (Koren, Bell, and Volinsky 2009). However, if we could find some currently active users who are predicted to be similar to these inactive users, then we can recommend videos based on active users' interests to get over the cold start problem.

5.1 Experiment Settings

From our dataset, we randomly selected two thousand active users (as ground truth) on the target day. To restrict the scope of neighbor selection without global searching from millions of users, we randomly drew five thousand candidate neighbors for each target user. For fair and unified comparison, we find top-K similar users from the candidate neighbors of each target user for recommending N videos by selecting the top-N popular videos among these K neighbors.

For neighbor selection in both tag-based (PTP & RTP) and video-based profiling (VBP) schemes, we utilized the similarity values predicted by the regression tasks in Sec. 4. For comparison, we also implemented various algorithms corresponding to different strategies (both personalized and non-personalized) of closest neighbor selection.

- Demographic profile similarity: select K closest neighbors according to the similarity of demographic profiles.
- Social friend filtering: select top-K close friends of the target user based on their interaction frequency, i.e., the monthly messaging days.
- Past long-term profiling: choose the top-K similar neighbors by comparing the past one month's video records of the target user and each candidate neighbor.
- Random: randomly select K users from the candidate neighbors.
- Global popularity: for each target user, always recommend the top-N popular videos among all the five thousand candidate neighbors.

5.2 Evaluation Metrics

Of each algorithm, we evaluate both the accuracy and diversity of the recommendation results. We utilize **F-measure** (Powers 2011) as the accuracy metrics which examines whether videos viewed by target users are ranked at top positions in the recommendation lists. Moreover, we use **diversification** as the diversity metrics which is defined as the average inter-user difference of recommendation results (Zhang, Zhou, and Zhang 2010). The larger this value, the more diverse the results are.

5.3 Experiment Results

Accuracy. Since the number of videos viewed by each user varies, we tested the performance by different values of N. Typically, the length for the recommendation list is in tens. Thus, our experimental study focuses on the interval [10,100]. We firstly fixed the value of K (= 15), and varied N[6]. Furthermore, one way to illustrate how different algorithms perform with regard to different number of neighbors, i.e, K, is to assume we know the number of videos viewed by each target user, which is equivalent to fixing N, and then vary the value of K. The results are in (Yang, Zhou, and Chiu 2016).

As shown in the results, PTP and RTP could more accurately hit users' interests than VBP, which validated the advantage of the tag-based scheme over the video-based profiling scheme. Global popularity based recommendation could achieve moderate performance because it is derived from the statistics and popular videos are just what the majority of users have viewed. Moreover, the reason why the social friend based strategy did not perform well as expected is that there are fairly few friends as candidates compared to other cases. In other words, even interest similarity between friends is in general larger than that between random users, friends may not be the most similar users to the target users among all the users.

Diversity. To further compare the diversity of recommendation results from tag-based and video-based profiling schemes, we tested the diversity for different values of N and K. The results can be seen in (Yang, Zhou, and Chiu 2016). Compared with recommendations from VBP, the results produced by two tag-based methods are more diverse in terms of inter-user difference. Another conclusion is that, although PTP is better than RTP with regard to prediction accuracy, RTP scheme can generate more diverse results, which is useful for discovery of the long-tail part of user interests.

6 Related Works

With the popularity of OSNs, a better understanding of how much two individuals are alike in their interests, namely, interest similarity, will benefit various applications in OSNs, such as friend recommendation (Lewis, Gonzalez, and Kaufman 2012), targeted online advertising (Yu and Houg 2014). In this paper, we apply the inferred interest similarity to video recommendation.

When we do not have users' behavioral data, such as for new users, users' interests can only be inferred based other user information, such as social cues which deduces a user's interests by considering this user's social neighbors' interests. Also interests could be inferred from the users who share more demographic attributes (Koren, Bell, and Volinsky 2009). A similar previous work investigated how various user information affects interest similarity based on video-based user profiling (Han et al. 2014). In this work, we mine and learn to predict two users' interest similarity with two tag-based user profiling methods, namely, PTP and RTP, which are shown to be more reasonable and effective than the traditional video-based method.

[6]We obtained similar results for other values of K.

7 Conclusion

In this paper, we mined and predicted users' interest similarity defined by tag-based interest profiles. By systematically studying the correlation between interest similarity and various user information, we select the most effective demographic, social and interest features. Then we test and compare the effectiveness of different feature combinations and models in predicting two users' interest similarity when their recent interests is few or blank. Furthermore, we apply our model to video recommendation to demonstrate its practical value. In the future work, we will try to explore and test more application scenarios for the interest similarity prediction.

Acknowledgement

The authors wish to acknowledge the support from HK RGC grant 14201814. This work was partially supported by the Natural Science Foundation of China under Grant No. 61402297.

References

Han, X.; Wang, L.; Park, S.; Cuevas, A.; and Crespi, N. 2014. Alike people, alike interests? a large-scale study on interest similarity in social networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, 491–496. IEEE.

Koren, Y.; Bell, R.; and Volinsky, C. 2009. Matrix factorization techniques for recommender systems. *Computer* (8):30–37.

Lewis, K.; Gonzalez, M.; and Kaufman, J. 2012. Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences* 109(1):68–72.

Powers, D. M. 2011. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.

Su, X., and Khoshgoftaar, T. M. 2009. A survey of collaborative filtering techniques. *Advances in artificial intelligence* 2009:4.

Yang, C.; Zhou, Y.; and Chiu, D. M. 2016. Who are Like-minded: Mining User Interest Similarity in Online Social Networks. arXiv:1603.02175v1.

Ye, J.; Chow, J.-H.; Chen, J.; and Zheng, Z. 2009. Stochastic gradient boosted distributed decision trees. In *Proceedings of the 18th ACM conference on Information and knowledge management*, 2061–2064. ACM.

Yu, D., and Houg, A. 2014. Facebook analytics, advertising, and marketing. In *Facebook Nation*. Springer. 117–138.

Zhang, Z.-K.; Zhou, T.; and Zhang, Y.-C. 2010. Personalized recommendation via integrated diffusion on user–item–tag tripartite graphs. *Physica A: Statistical Mechanics and its Applications* 389(1):179–186.