Enhancing Collaborative Filtering with Friendship Information

Liliana Ardissono, Maurizio Ferrero, Giovanna Petrone, Marino Segnan
Dipartimento di Informatica, Università di Torino
Corso Svizzera 185
Torino, Italy 10149
[liliana.ardissono,giovanna.petrone,marino.segnan]@unito.it,maurizio.ferrero@edu.unito.it

ABSTRACT

We test the impact of integrating a measure of *common friendship* in collaborative filtering, in order to capture the intuition that socially interconnected groups of people tend to have similar tastes. An experiment on the Yelp dataset shows that using preference information derived from the commonalities of interests in networks of friends achieves higher accuracy than item-to-item collaborative filtering.

CCS CONCEPTS

•Information systems → Recommender systems; •Human-centered computing → Collaborative Filtering; Social recommendation;

KEYWORDS

Recommender systems; homophily and social networks

1 INTRODUCTION

Several models have been proposed to integrate explicit and implicit social information in collaborative recommenders; e.g., [2, 6]. We are interested in analyzing the impact of "group-based" friendship relations, which social science has associated to user similarity through the homophily phenomenon, according to which "similarity breeds connection" [7]. As homophily has been observed in several types of social networks, including digital ones [1], it is worth studying its relevance to collaborative recommender systems, which employ rating and/or tagging similarity for item suggestion.

Our research questions are: *"RQ1: Can the performance of a collaborative filtering recommender be improved by taking into account common friendship relations in groups of people? RQ2: How does this type of information influence performance if taken alone, or in combination with other sources of data about the user, such as rating behavior?"* In order to answer these questions, we compared the performance of collaborative filtering recommenders based on user ratings, community membership, group-based friendship relations, product selection, and combinations of these types of information. We tested the recommenders on a subset of the Yelp dataset [8] providing data about friends relations and item ratings on restaurants. The experiment showed that the integration of social and rating information outperforms the other algorithms, even though, as observed in previous works, there is a trade-off between recommendation accuracy and coverage.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
UMAP '17, July 09-12, 2017, Bratislava, Slovakia
© 2017 Copyright held by the owner/author(s). ACM ISBN 978-1-4503-4635-1/17/07.
DOI: http://dx.doi.org/10.1145/3079628.3079629

2 INTEGRATING SOCIAL INFORMATION IN COLLABORATIVE FILTERING

2.1 Dataset

The Yelp dataset [8] provides information about friendship relations and user ratings on various types of businesses. We selected the data about restaurants, considering only the users who rated at least 20 items. This restricted dataset (henceforth, "Yelp-Restaurants(20)") includes 8914 users, 23210 items and 419013 ratings. Its user-rating matrix has sparsity = 0.9979. An analysis of the structure of the social network underlying the dataset shows that: (i) The cumulative distribution of the number of ratings in the observed population follows the Power Law, with most users having rated few items. (ii) The cumulative distribution of the number of friends per user follows the Power Law, with most users having few or no friends. (iii) There is a positive correlation between the number of ratings provided by users and the number of friends they have: the most "isolated" users rated few items.

2.2 Recommendation Algorithms

We describe the Collaborative Filtering (CF) algorithms that we tested on the Yelp-Restaurants(20) dataset. All the algorithms are based on the K-nearest neighbors approach with K = 10. In the description we adopt the following notation: u is the user for whom the predictions are computed and v is another user; i is the item for which u's rating is estimated and j is another item; $\hat{r}_{ui}$ is the estimate of u's rating of i; $\overline{r}_i$ ($\overline{r}_j$) is the average rating received by item i (j); $\overline{r}_u$ ($\overline{r}_v$) is the average rating given to items by user u (v).

Item-to-item CF.

This algorithm assumes that u's preference for i can be inferred from the ratings (s)he gave to items j that the other users rated similarly to i; $\hat{r}_{ui}$ is computed as follows:

$$\hat{r}_{ui} = \overline{r}_i + \frac{\sum_{j \in N_u(i)} \sigma(i,j)(r_{uj} - \overline{r_j})}{\sum_{j \in N_u(i)} |\sigma(i,j)|} \quad (1)$$

where: $N_u(i)$ is the set of neighbor items of i that have been rated by u; $\sigma(i,j)$ is the Pearson Similarity between i and j tuned with significance weighting; see [3]. $\sigma(i,j)$ is used to identify i's neighbors[1] and to weight their contributions to $\hat{r}_{ui}$.

Community-based user-to-user (U2U) CF.

This algorithm uses the ratings provided by u's neighbors from her/his community to estimate u's ones. As an aggregation factor for the formation of communities we considered direct friendship

[1] In order to select very similar items, only those for which $\sigma(i, j) \geq 0.5$ are considered as candidate neighbors of i.

relations *and* the number of common friends among users. Given u's community of friends, $\hat{r}_{ui}$ is computed as follows:

$$\hat{r}_{ui} = \overline{r}_u + \frac{\sum_{v \in N_i(u)} \sigma(u,v)(r_{vi} - \overline{r}_v)}{\sum_{v \in N_i(u)} |\sigma(u,v)|} \quad (2)$$

where: $N_i(u)$ is the set of neighbors of u (according to $\sigma(u,v)$) who rated i. Moreover, $\sigma(u,v)$ is the modified Jaccard Similarity ($JS'(u,v)$) between u and v, computed by taking the number of common friends, and direct friendship relations, into account:

$$\sigma(u,v) = JS'(u,v) = \frac{|Friends_u \cap Friends_v| + 1}{|Friends_u \cup Friends_v|} \quad (3)$$

$JS'(u,v)$ captures the concept of direct and mutual friendship in user groups: when u and v have no common friends, $JS'(u,v)$ returns a positive value so that they are not excluded from the set of candidates for preference estimation. However, it returns higher values for people socially connected at the group level.

Tag-based U2U CF.

The idea is that, as tags describe item types, rating an item j provides evidence of interest in other items having tags in common with j; e.g., see [4]. Moreover, the number of ratings given by the user to items tagged as t provides evidence about her/his degree of interest in t. Therefore, u's interests can be described as a vector X_u specifying the number of occurrences (frequencies) of the tags associated to the items (s)he rated: $X_u =< freq_{t1}, \ldots, freq_{tn} >$. Preferences are estimated user-to-user, by means of Equation 2. However, user similarity (denoted as $\sigma_T(u,v)$) is based on the common tags occurring in u and v's vectors. Specifically, $N_i(u)$ is the set of neighbors v of u who rated i, with $\sigma_T(u,v) \geq 0.5$ to select very similar neighbors. Moreover, $\sigma_T(u,v)$ is the cosine similarity between X_u and X_v, modified by means of significance weighting to consider the number of common tags in the two vectors:

$$\sigma_T(u,v) = \frac{\min_(|T_{uv}|, \gamma)}{\gamma} cosineSimilarity(X_u, X_v) \quad (4)$$

T_{uv} is the number of tags occurring in both X_u and X_v, and γ is threashold set to optimize the accuracy (F1) of the algorithm via regression testing.

Community+Tag-based U2U CF. We consider the joint contribution of community and tag-based similarities to generate predictions, assuming that they corroborate each other. Here, $\hat{r}_{ui}$ is computed using Equation 2, applied to users belonging to u's friends community. However, the similarity between users is the sum of the modified Jaccard Similarity (Equation 3) and the Tag-based one (Equation 4), normalized in [0, 1]: $\sigma(u,v) = \frac{JS'(u,v)+\sigma_T(u,v))}{2}$

Friends-based U2U CF.

Given the principle that similarity breeds connection, and the results reported in [1], this algorithm exploits a selection of u's neighbors from her/his direct friends, considering the number of common friends among them, to predict u's preferences. Here, $\hat{r}_{ui}$ is computed using Equation 2, where $\sigma(u,v) = JS'(u,v)$ is the modified Jaccard Similarity between u and v; see Equation 3.

FilteredFriends-based U2U CF.

This algorithm exploits both friendship and rating behavior to estimate preferences, using these factors to select u's neighbors. We defined a hybrid user-to-user recommender that works as follows: firstly, it selects a set of candidate neighbors of u based on the rating similarity on the items having at least one tag in common with i. Then, it sorts the set of candidates by friendship similarity, using the modified Jaccard Similarity in the network of u's direct friends, and it selects the best K neighbors on the basis of $JS'(u,v)$. We considered two similarity thresholds for the selection of neighbors to investigate the impact of social proximity and connection on prediction capabilities: $JS'(u,v) \geq 0$ and $JS'(u,v) \geq 0.1$. Then, it computes $\hat{r}_{ui}$ using Equation 2, where $\sigma(u,v)$ is $JS'(u,v)$.

3 EXPERIMENTAL RESULTS

We tested the performance of the previous algorithms on the Yelp-Restaurants(20) dataset by applying 5-fold cross-validation, after having randomly distributed users on folders. For each user, we used 80% of the ratings as learning set and 20% as test set. We evaluated the recommenders on their best 10 predictions. The evaluation produced the following results:

1) The Community-based U2U CF recommender obtains poor accuracy values, confirming that communities fail to provide specific information about user preferences [5].

2) Friends-based U2U CF ($JS'(u,v) \geq 0$) obtains poor accuracy values. However, by restricting neighbors to direct friends who have several common friends ($JS'(u,v) \geq 0.1$), it achieves the third-best accuracy results.

3) The combination of friendship and rating-based similarity is the most promising approach regarding accuracy. FilteredFriends U2U CF ($JS'(u,v) \geq 0.1$) outperforms all the other algorithms by restricting the pool of candidate neighbors to common friendship. However, it has limited user coverage (41%).

4 CONCLUSIONS

We compared the performance of different Collaborative Filtering recommenders to evaluate the usefulness of integrating social information with data about rating behavior. We discovered that the accuracy of predictions can improve by restricting the set of neighbor users to those belonging to the network of highly interconnected friends whose interests are similar to the user's ones.

This work was funded by the University of Torino in the "Ricerca Locale" support program.

REFERENCES

[1] L.M. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines, and F. Menczer. 2012. Friendship prediction and homophily in social media. *ACM Transactions on the Web (TWEB)* 6, 2 (2012), art. 9.

[2] A. Bellogín, I. Cantador, and P. Castells. 2010. A Study of Heterogeneity in Recommendations for a Social Music Service. In *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems*. ACM, New York, NY, USA, 1–8.

[3] C. Desrosiers and G. Karypis. 2011. A Comprehensive Survey of Neighborhood-based Recommendation Methods. In *Recommender systems handbook*, F. Ricci, L. Rokach, B. Shapira, and P.B. Kantor (Eds.). Springer, 107–144.

[4] J. Gemmel, T. Schimoler, B. Mobasher, and R. Burke. 2012. Resource recommendation in social annotation systems: a linear-weighted hybrid approach. *Journal of computer and system sciences* 78 (2012), 1160–1174.

[5] X. Xu H. Bisgin, N. Agarwal. 2010. Investigating Homophily in Online Social Networks. In *Proc. of Web Intelligence and Intelligent Agent Technology (WI-IAT)*. Toronto, Ontario, Canada, 533–536.

[6] X. Liu and K. Aberer. 2013. SoCo: A Social Network Aided Context-aware Recommender System. In *Proceedings of the 22Nd International Conference on World Wide Web*. ACM, New York, NY, USA, 781–802.

[7] M. McPherson, L. Smith-Lovin, and J. Cook. 2001. Birds of a feather: homophily in social networks. *Annual review of sociology* 27 (2001), 415–444.

[8] Yelp. Yelp Dataset Challenge. https://www.yelp.com/dataset_challenge.