Social Network Collaborative Filtering

Rong Zheng, Foster Provost, Anindya Ghose

Abstract

This paper reports on a preliminary empirical study comparing methods for collaborative filtering (CF) using explicit data on consumers' social networks. To our knowledge it is the first study to carefully evaluate the potential of explicit, publicly represented social networks for making product recommendations. Understanding social-network CF is important because traditional CF over a large consumer base is tremendously expensive computationally. An often-ignored aspect of CF is the selection of the set of users from which to make recommendations. Social theory tells us that social relationships are likely to connect similar people. If this similarity is in line with the recommendation task, they may provide a small, dense set of "recommenders" for CF. We examine a unique dataset from Amazon.com that contains a social network of consumer-selected friends. We examine two ways to incorporate social-network information into CF: using the social network to restrict the set of recommenders selected, and (further) using proximity in the social network to modify the traditional CF calculation. The results show that that CF with social-network members selected as recommenders can be remarkably superior as compared to collaborative filtering with the recommenders not socially connected. Once the social network is selected, social network proximity does not seem to improve recommendations.

1. **Introduction**

Explicitly represented, user-generated social networks are becoming increasingly important in modern technology-mediated consumer interaction. Consumers have flocked to social-networking sites, revealing various information about themselves, and connecting themselves to others. One important potential ramification of the increased availability and accessibility of information about consumers is that it may be useful to help users to find items that they would like to purchase. This has obvious implications for e-commerce firms, such as Amazon.com, who can increase sales via product recommendations. Additionally, users themselves may be able to find products better by examining the preferences implicit or explicit in the information revealed by prior purchasers of a product (e.g., via product reviews) or by their social-network neighbors.

To our knowledge, prior research has not evaluated the potential of explicit social networks[1] for making product recommendations. Off-line social networks have been the subject of intense academic research for many decades, but studies have been limited by the difficulty of gathering either social links or purchase information or both (Hill et al., 2006). An exception, Hill et al. (2006) examined a large data set including both a communication-based social network and specific information on a targeted marketing campaign. They showed that the social network could be used to great advantage to determine the consumers likely to purchase, even in the presence of sophisticated modeling using consumer-specific information. However, their setting did not include a variety of different products, and so was not amenable to traditional recommendation techniques. (They review the literature on the use of social networks for marketing and research, to which we refer the interested reader.)

---

[1] For the sake of clarity and focus, in this paper we distinguish *social* networks based on actual friend or acquaintance links, from induced consumer relationship networks based on co-purchase or other non-social relationships. Some prior work (e.g., Perugini et al. 2004; Huang et al. 2004) has noted that some recommendation systems induce an indirect interaction network over consumers.

Recommender systems generate product recommendations in order to reduce consumers' search costs in light of the increasing product variety on the Internet (Resnick and Varian 1997). Using past information about consumers and products, these systems identify promising future interactions between consumers and products, and present to users information about items they are most likely to be interested in. The goal of a commercial recommender system includes both increasing sales and increasing user satisfaction. This paper examines *collaborative filtering* (CF), perhaps the most well-known recommendation technique, which has been used widely in e-commerce applications and especially in academic research (Adomavicius and Tuzhilin 2005). The underlying principle behind "user-based" CF is to find customers who purchased the same or similar items. Importantly, CF techniques are very expensive computationally, having to compare large numbers of users to each other—which is one reason why Amazon.com instead uses alternative recommendation techniques (Linden et al., 2003).[2]

This paper reports on an empirical study comparing CF techniques that incorporate information on consumers' social networks. Social theory tells us that social relationships are likely to connect similar individuals (McPherson et al., 2001). Therefore, incorporating social network information into CF may be beneficial: it may be able to restrict the users considered by CF to individuals likely to provide useful information, and thereby avoid tremendous computational expense. We examine a unique dataset from Amazon.com that contains both a network of consumer-selected friends and recent product purchases. We introduce two possible components of social-network-based collaborative filtering (SNCF): using the social network to restrict the universe from which recommenders are selected, and using proximity in the social network to modify the ordering of recommendations. The results show that that CF with social-network members selected as recommenders can be remarkably superior as compared to collaborative filtering with the recommenders not socially connected. Once the social network is selected as the recommender set, social network proximity does not seem to improve recommendations.

2. Social-Network Collaborative Filtering (SNCF)

In the typical setting, commercial collaborative filtering exploits the interaction data between consumers and products and makes predictions of products a consumer will purchase. The input of the problem is an $M \times N$ *interaction matrix* $T = (t_{ij})$ associated with M consumers $C = \{c_1, c_2, \ldots, c_M\}$ and N products $P = \{p_1, p_2, \ldots, p_N\}$. We focus on recommendation that is based on transactional data (rather than rating data). That is, a_{ij} can take the value of either 0 or 1 with 1 representing an observed transaction between c_i and p_j and 0 absence of transaction. *User-based* collaborative filtering algorithms first compute a consumer similarity matrix $W = (w_{st})$, s, t =1, 2, …, M. The similarity score w_{st} is calculated based on the row vectors of A using a vector similarity function. A high similarity score w_{st} indicates that consumers s and t may have similar preferences since they have previously purchased a large set of common products. $W \cdot A$ gives potential scores of the products for each consumer.

For a given target user, consider as the set of "recommenders" the user-base to whom the target will be compared to generate recommendations. The first component to SNCF is to choose the set of potential recommenders used by CF as those in the target user's social network, with the hope that the social network will provide a dense set of users with similar tastes.[3]

---

[2] Consider that Amazon.com has 50 million or more active customers; a single "long-tail" category like books can have over a million items.

[3] How practicable this is depends on both the user and the firm. Tens of thousands of users belong to Amazon's friends network, and Amazon knows their purchase history. Orders of magnitude more belong to social networking sites, and

An explicit consumer social network can be represented as a graph with nodes being the consumers and links being the social relationships among them. The second potential component of SNCF is to modify the selection or ranking of recommendation based on the structure of the social network. For this study, we introduce a straightforward modification. Specifically, instead of using similarity between past purchasing behavior to find consumers with similar purchasing preferences, *proximity-based SNCF* uses the distance between consumers in the social network. We adopt the standard graph-theoretic definition of distance of nodes: the minimal number of edges that link the nodes. Therefore the first step of the similarity computation is to find the minimum number of edges between two nodes. The input is a graph which is represented by an adjacent square matrix $G = (g_{st})$, s, t =1, 2, …, M; g_{st} can be 1 or 0 depending on if there is an edge between consumer s and consumer t. The matrix G is symmetric if the links on the graph have no direction. The output is a distance matrix t $D = (d_{st})$ s, t =1, 2, …, M . These distances can by computed by Dijkstra's algorithm, or when the social-network links are unweighted, simply via a breadth-first search (which is the case for this study). Then, under the assumption that social influence will decay exponentially as the social-network distance increases, the distance matrix is transformed to the *influence* matrix $I = (i_{st})$ s, t =1, 2, …, M via: $i_{st} = \exp(-d_{st})$. In direct analogy to collaborative filtering, the community-based scores for the potential recommendations are calculated by $I{\cdot}A$. Below we will create different "versions" of the algorithm by limiting the span of influence to $d \leq k$, for particular values of k. So, for example, we can look at the influence only of direct neighbors by setting k=1.

3. **Data and Experimental Setup**

We generate recommendations for a subset of 1206 Amazon.com customers who have chosen to reveal their purchases on Amazon's site. The total set of purchases includes all revealed purchases made by these consumers over the three-month period between May and July 2007. In sum, a total of 11,773 distinct items were bought by these 1206 consumers. About 50% of the purchased items are books; 40% are CDs and DVDs, and the remaining 10% include products from other categories such as electronics, apparel etc. Importantly for this study, one-half of these 1206 customers are interconnected by the "Amazon Friends" social network.[4] The set of 603 purchase revealers who have at least one "friend" who also has revealed her purchases, will be the social network chosen as recommenders and as targets for SNCF.

We divide the purchases by timestamp. The 20% most recent purchases for each consumer are held out for prediction; the 80% older purchases will be used to make recommendations. Each recommender system calculates a score for each potential user/product pair, resulting in a ranked list of recommendations. For this study, we use standard precision/recall analysis to evaluate the quality of this ranked list of recommendations, examining the possible tradeoffs between the accuracy of recommendations (precision) and coverage of actual purchases (recall). More specifically, for any number t of desired recommendations, the top-t recommendations (user/item pairs) are chosen and precision and recall are calculated in the usual manner. We should keep in mind that, in contrast to evaluations where users rate the desirability of all products, for this "actual purchase" prediction we do not expect very high precision or recall; predicting actual purchases is a very hard problem.

---

some firms have exclusive marketing contracts with them. Furthermore, Google, Microsoft, eBay, AT&T, Verizon, and many other firms have access to social networks via phone calls, Skype calls, IM messages, email, and so on.

[4] The 1206 customers are the intersection of the purchase-revealers and the Amazon Friends network; one-half have social-network neighbors who are also purchase revealers.

4. Results

Our first experiment assesses the two SNCF components comparatively, assuming the social-network has been selected as the recommender base. Then we expand to the whole data set. Figure 1 plots precision/recall curves for "purchase-based" SNCF (using the social network to select the user base and regular CF for recommending), proximity-based SNCF,[5] and a hybrid of the two where the non-zero purchase-based scores are boosted in a linear combination with the proximity-based scores.[6] Clearly, the proximity-based adjustments result in inferior recommendations. Perhaps this should not be surprising: traditional, purchase-based CF is designed specifically for making recommendations from consumers with similar tastes. As shown by the third curve, we were not able to build a hybrid of the two that improved significantly, and the small size of the data set makes it imprudent to try too hard, lest the process overfit the data set.

In an absolute sense, the recommendation accuracy here is remarkable. The precision for the top-100 recommendations is around 20 percent, and for the top-1000 recommendations is still around 10 percent—with a recall of 5%. Previous studies of actual-purchase-based recommendations do not come close to these precision/recall tradeoffs (Huang et al., forthcoming).[7]

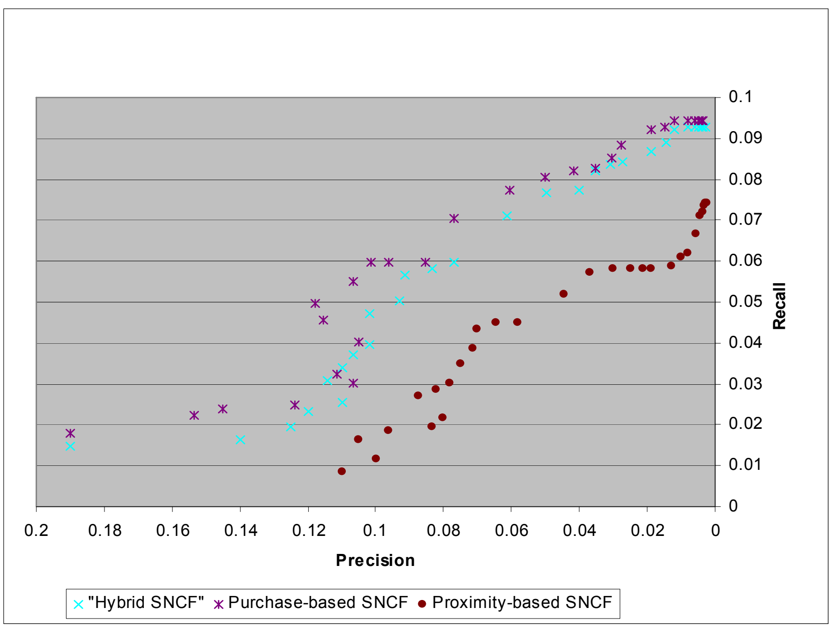

Figure 1. Precision/recall comparison of three Social-Network Collaborative Filtering techniques. Purchase-based uses the traditional CF calculation limited to the users in the social network. Proximity-based additionally finds similar users based on distance in the social network. Hybrid tries to boost the purchase-based scores using the proximity scores.

For comparison, Figure 2 shows the performance of CF without the social-network user selection (using the other 603 customers, who are not part of the social network). The precision and recall results are an order of magnitude worse than for SNCF—demonstrating just how much advantage is conferred by the selection of the social network recommenders. For data sets of this size at least, CF without social-network selection just can't compare with SNCF. Note that except for the fact that they are not networked socially, these other 603 customers are similar to those in the social network (e.g., in terms of number and variety of purchases).

---

[5] The proximity-based SNCF in Figure 1 uses k=2; the results are similar for larger k. Using k=1 does not produce enough recommendations for competitive recall (early-curve precision is slightly better).

[6] Here, the combination is purchase_score + 0.5 proximity_score. We tried to learn a good combination weighting; however, apparently there are too few positives for effective cross-validated training.

[7] The numbers are not completely comparable, but the implication is clear.

When CF is applied to the union of the two data sets—the results still are substantially worse than those shown in Figure 1: the high-precision from the intra-network predictions still is evident, but for a given level of precision recall is cut in half (because only the social-network recommendations play any considerable role in the accuracy). What's worse, the running time increases super-linearly in the size of the user-base (Linden et al. 2003). Generally one would run CF on much more than just 1206 users.

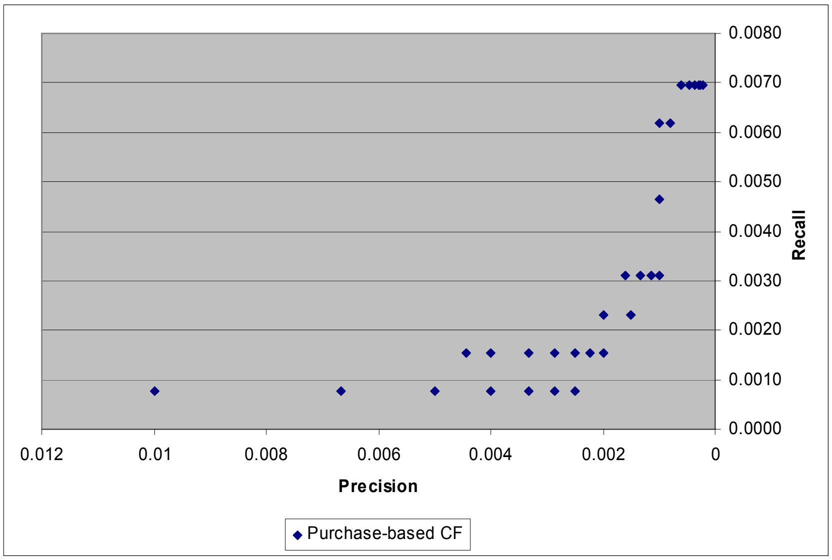

Figure 2. Traditional purchase-based CF on a data set almost identical in size and composition to that used for the SNCF results presented in Figure 1, but where the users do not form a social network. Note the order-of-magnitude difference in scale on both axes.

5. **Other Related Literature**

Large real-world networks such as the world-wide web, internet topology, social networks, biological networks, and linguistic networks have been extensively studied from a structural point of view. Typically, these studies address properties of the graph including its size, density, degree distributions, average distance, small-world phenomenon, clustering coefficient, connected components, community structures, etc. (Nowelll et al. 2005). Online friendship and email graphs have been studied in the context of explaining and analyzing friendships (Kumar et al. 2004) and demonstrating the small-world and navigability properties of these graphs (Dodds et al 2003, Nowell et al. 2005, Adamic and Adar 2005). However, none of this work has examined the impact of social network based relationships on members' affinity to purchase products, particularly with the objective of designing recommender systems.

An emerging stream of literature in computer science and marketing has also analyzed the efficacy of recommender systems. Prior work in recommender systems has postulated that "recommendations, however, are not delivered within a vacuum, but rather cast within an informal community of users and social context" (Perugini et al. 2004). Recent research (Huang et al. forthcoming, Miraza et al. 2003) improved the quality of recommendations by extending the direct co-purchase relationship to an indirect co-purchase network. A limitation of this stream of work is that the co-purchase behavior only accounts for one kind of relationship among consumers—that based on product purchases. As a consequence, if product preferences do not represent the complete antecedents of buyer purchases, the predictive performance of this kind of collaborative filtering method will be affected. Our work contributes to this stream of research by demonstrating the impact of social information on product purchases. Finally, the implications of our work are related to the emerging stream of work on word-of-mouth that captures how the inherent trust embedded in user opinions and social information disclosures in online communities affect product sales (Chevalier and Mayzlin, forthcoming).

5. Discussion

Clearly these results are based on a single study, on a relatively small data set—and should be taken as a preliminary study. Nevertheless, they show convincingly that recommendations made by social-network-based collaborative filtering can be far superior to recommendations made by collaborative filtering on a similar-size data set that does not represent a social network. These results can be seen in (at least) two ways. First, they add support to the results of Hill et al. (2006) that social networks can enable technology-based methods to predict purchase behavior, and to our knowledge this is the first study to show the effect in a collaborative filtering setting. Second, they demonstrate a very effective way to scale up collaborative filtering, a technique previously thought to be inapplicable to large user bases (Linden et al., 2003). Specifically, CF can be scaled up by using social networks to scale down the user base used to make recommendations to a particular user. In our experiments, doubling the size of the user base did not improve the precision of the recommendations; recall was cut in half, and computational cost more than doubled (because the number of users and the number of products each double). Furthermore, the results on the larger user base only look good because the SNCF results are embedded—separated out, the non-social network results are an order of magnitude worse.

References

Adamic, L., and E. Adar. How to search a social network. Social Networks, 27(3):187–203, 2005.

Adomavicius, G., and A. Tuzhilin, Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions, IEEE Trans. Knowl. Data Eng. 17(6): 734-749, 2005.

Breese, J.S., Heckerman, D. and Kadie, C., Empirical analysis of predictive algorithms for collaborative filtering. in Fourteenth Conference on Uncertainty in Artificial Intelligence, (Madison, WI, 1998), Morgan Kaufmann., 43-52.

Chevalier, J. and Mayzlin, D., "The Effect of Word of Mouth on Sales: Online Book Reviews" Journal of Market Research, forthcoming.

Dodds, P., R. Muhamad, and D. J. Watts. An experimental study of search in global social networks. Science, 301:827–829, 2003.

Hill, S., F. Provost, and C. Volinsky. "Network-based Marketing: Identifying likely adopters via consumer networks. " Statistical Science 21 (2) 256–276, 2006.

Huang, Z., H. Chen, D. Zeng, "Applying Associative Retrieval Techniques to Alleviate the Sparsity Problem in Collaborative Filtering. ACM Transactions on Information Systems (TOIS), 22(1), 116–142 (2004).

Huang, Z., D. Zeng, H. Chen, "A Comparative Study of Recommendation Algorithms in E-commerce Applications," IEEE Intelligent Systems, forthcoming.

Kumar, R., J. Novak, P. Raghavan, and A. Tomkins. Structure and evolution of blogspace. CACM, 47(12):35–39, 2004.

Liben-Nowell, D., J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. PNAS, 102(33):11623–11628, 2005.

Linden, G., B. Smith, and J. York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, page 7680, 2003.

McPherson, M., L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.

Mirza, B.J.,Keller, B.J., and Ramakrishnan, N. (2003). Studying Recommendation Algorithms by Graph Analysis. Journal of Intelligent Information Systems, 20(2), 131–160.

Resnick, P. and Varian, H. Recommender systems. Communications of the ACM, 40 (3). 56-58, 1997.

Perugini, S., Goncalves, M. A., & Fox, E. A. (2004). A connection centric survey of recommender system research. Journal of Intelligent Information Systems 23[1].