Does Trust Influence Information Similarity?

Danielle H. Lee & Peter Brusilovsky
School of Information Sciences
University of Pittsburgh
135 N. Bellefield Ave. Pittsburgh, PA, USA
+1-412-624-9437
{hyl12, peterb}@pitt.edu

ABSTRACT
In collaborative filtering recommender systems, users cannot get involved in the choice of their peer group. It leaves users defenseless against various spamming or "shilling" attacks. Other social Web-based systems, however, allow users to self-select trustworthy peers and build a network of trust. We argue that users self-defined networks of trust could be valuable to increase the quality of recommendation in CF systems. To prove the feasibility of this idea we examined how similar are interests of users connected by a self-defined relationship in a social Web system, *CiteuLike*. Interest similarity was measured by similarity of items and meta-data they share. Our study shows that users connected by a network of trust exhibit significantly higher similarity on items and meta-data than non-connected users. This similarity is highest for directly connected users and decreases with the increase of distance between users.

Categories and Subject Descriptors
H.1.2 [**User/Machine Systems**]: Human Factors; Software Psychology; J.4 [**Social and Behavioral Sciences**]: Sociology

General Terms
Measurement, Human Factors

Keywords
User Similarity, Trust, Human Network

1. INTRODUCTION
Recommender systems powered by collaborative filtering (CF) technologies become a feature of our life. Such popular systems as Amazon.com, Netflix, Last.fm, and Google News use Collaborative filtering to recommend us products to buy, movies to watch, music to listen and news to read. The power of this technology is based on a relatively simple idea: starting with a target user's rating, find a peer cohort (neighborhood) of users who have similar interests and recommend items favored by this cohort to the target user. As such, the choice of cohort is an essential part in CF recommendations and is usually determined by automatically calculating rating similarities between the target user and other users. In a typical CF system, this peer cohort (a group of users selected as the basis for CF) is unknown to target users. Moreover, the target users cannot add trustworthy users to their cohort group nor exclude suspicious users from the group.

The success of social linking and bookmaking systems that allow users to build their networks of trust, stresses a fact forgotten by modern CF systems: the source of the recommendation is an important criterion for judging the quality of recommendations [2]. A range of Web 2.0 systems such as *LlinkedIn*, *Flickr*, *Delicious*, *CiteuLike* etc., provide various kind of social linking, enabling their user to pick known and trusted users and add them to their list of connections. These self-defined links between users establish a rich network of trust, which is, in turn, used to propagate various kinds of information. Given that, it is natural to expect some kind of merger between social linking and CF technology: a new generation of *trust-based recommender systems*, which will use self-defined social networks of trust to improve the quality of CF systems and the satisfaction of their users. Some pioneer works in this direction already appeared [4, 5, 6, 9, 13]

To prove that trust-based recommenders are more than a speculation, some important assumptions have to be checked. Is it true that connected users in the networks of trust share not only trust, but also some common interests? Is it true that information can flow along these networks, i.e., the choices made by users are affected by the choices of users they trust? The goal of this paper is to test these assumptions. Using real life data collected from a social Web system, *CiteuLike*, we examined several important properties of self-defined trust networks. We investigated how similar are users' interests in these networks, the extent to which amount of similar information collected by users depends of the strength of their connection, and ultimately, how feasible it may be to use a network of trust for personalized recommendation.

The term 'trust' as used in this paper may not be an exact match with the general use of 'trust' as defined in the sociology. The social relationship used in this paper is defined unilaterally, simply indicating user trust in the usefulness of information provided by connected individual. It is not trust through personal interaction or emotional support (for instance, connected with an expectation of obligation, morality or responsibility [7]). Since referred users are deemed "trustworthy" by the target user in terms of information collection, however, the term 'trust' was selected. Furthermore, the term 'trust' as defined in the Webster's Third New International Dictionary meets our interpretation of 'trust.' Its definitions for the term are "a confident dependence on the character, ability, strength, or truth of someone or something," "confident anticipation," and "a charge or duty imposed in faith and confidence or as a condition of some relationship" [7]. To date, a better or more precise term for this relationship has not been found; hence, *trust* is used hereafter.

2. RELATED WORK
The popularity of CF technology, revealed some problems. CF appeared to be not well-protected against malicious users who try to harm the system or to make a profit by gamming the system. For example, by copying the whole user profile, a malicious user is perceived by the system to be a perfect peer user and the products added by him are therefore recommended to the target user [3, 5, 8]. Even without malicious users the quality of recommendation can be affected by peculiar users with unusual

interests [10]. Moreover, since CF systems have to compare all other users in order to find the peer group, the computation requires substantial off-line process [4]. Finally, users who do not have sufficient ratings are not able to receive reliable recommendations [10]. These CF-related problems occur in part because the recommender systems make a choice of peer group purely by similarity computation, and do not allow the target users to affect this part of the recommendation process.

Several research teams attempted to exploit trust between users to resolve some of the cited problems of CF technology. Massa and Avesani's study [4] showed that a user's trust network can solve the ad-hoc user problem, improve recommendation prediction and attenuate the computational complexity. Another study indicated that a trusted network decreases the recommendation error and increases the accuracy as well [9]. For users with a unique taste, their own trusted network could increase the satisfaction of recommendation, since they are able to know where the information came from [12]. The recommendations made by friends were known to be frequently better and more useful than the recommendation made by systems [11].

To prove the feasibility of trust network as a source of information for reliable recommendation, several research teams started with checking the main assumption: do users linked by self-defined networks of trust have similar interests.

Singla and Richardson (2008) found the positive correlation of frequency and time of instant messaging between users with search interests [11]. Another trust-related research suggested that two users who are friends tend to share similar vocabularies, in-links and out-links on their personal homepages [1]. Ziegler and Golbeck [13] compared interest similarity between people in a trusted network. They used information regarding users and the user's trust ratings in the book recommendation. Rather than using each information item, they grouped the items by topics, using an existing taxonomy. Then, they built topic-based user profiles and the closeness of the user profiles in the trusted network was assessed. As the conclusion, they found that topic-based user profiles became more similar as the trust values between two users increased and reduced the data sparsity problem existing on the comparison of individual item [13]. Our work presented below was motivated the same goal: to assess interest similarity between users connected by relationship of trust.

3. DATA COLLECTION

3.1 Data Sets

As a source of data for our study we selected *CiteuLike*, a social Web system for sharing bibliographic references. To pick up initial set of users, we visited this site randomly in September and October of 2008. Users who posted new articles at the time of visit were picked. The information collected for each user included the bibliography (article title, list of authors, journal name, publication year, etc.) and the *watchlists* (connected users). After collecting a group of initial users, we collected data of their trusted connections. Table 1 shows the descriptive statistics.

In collaborative tagging systems explicit connections between users are of special nature. In some sense, they bear more "trust" than the connections between friends in social networking systems. In *CiteuLike*, users can directly connect to other users who have interesting bibliography by adding them '*watch list*.' Then the system displays the whole bibliography of watched users.

Table 1. Data Summary of *CiteuLike*

Total no. of users	21076
Total no. of distinct items (papers)	449824
Average no. of items per user	28.69
Total no. of unidirectional relation	11295
Total no. of reciprocal relation	93

3.2 The Networks of Trust

In this paper, we interpret user's act of connecting to other users (by adding this user to the watch list) as a sign that she likes the focus and trust the quality of the added user's references and wants to have direct access to them continuously in future. Thus, watching in *CiteuLike* could be considered as evidence that connected users are trustworthy to the original user in terms of information collection.

We distinguish two kinds of trusted connections – unidirectional and reciprocal. The act of adding another user to the '*watch list*' is unidirectional (which is different from social networking systems). If user A added user B to her network, it does not imply that user B will be added to A's network necessarily. The users in A's network decide independently whether to add A to their networks. For example, user B may not have A in his network and we call the relationship between A and B as 'unidirectional'. Another user C in A's network may add A to his network as well. We call this relationship as 'reciprocal' (Figure 1).

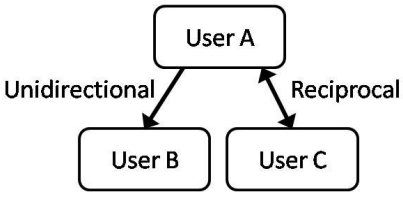

Figure 1. Directions of relation in the center of user A

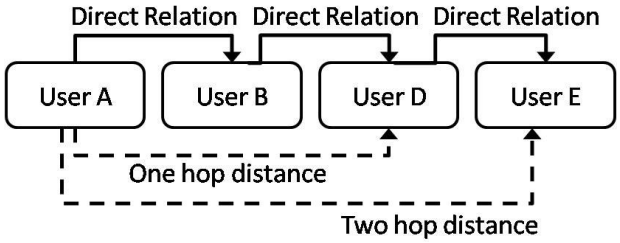

Figure 2. Relation distance in the center of user A

We also distinguish distances of connections to investigate the transitivity of common interests in the networks of trust. Three distances between users in trust networks were explored: *direct*, *one hop* and *two hops*. In the above example, user A and user B are in 'direct' relationship. If user B is trusting user D, user A and user D are in 'one hop distance' unidirectional relationship (Figure 2). If user E belongs to the watch list of user D, users A and E are in 'two hop distance' unidirectional relationship. These distances can be applied to the reciprocal relationship as well.

4. DATA ANALYSIS

In this study we tested how similar the information shared by people in trusted network is. Specifically, we counted the number

of shared information *items* (academic papers) and *meta-data*. In *CiteuLike* context, authors and journals (or conferences) is a good example of meta-data.. In our study, however, we considered authors only since it is more reliable and easy to track. Following Ziegler and Golbeck [13] experience with topics (which is another kind of metadata), we expected that the users who share the same interests may not necessarily agree about specific items, but demonstrate higher agreement on the level of meta-data (authors).

Since sizes of item collections varied dramatically from user to user, we had to examine both absolute and relative similarity measures. I.e., in order to measure between-users' information similarity, we not only used absolute numbers (i.e., number of common items), but we also compared *relative* (normalized) Jaccard similarity: proportion of shared items in respect to the whole collections of connected users. We used three meaningful relative similarity measures as dependent variables. Figure 3 and the following equations explain the meaning of these measures.

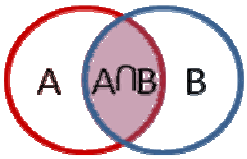

Figure 3. Information Overlap

$$\text{Inlink Power} = (A \cap B)/A \quad \text{eq. (1)}$$

$$\text{Outlink Power} = (A \cap B)/B \quad \text{eq. (2)}$$

$$\text{Overall Power} = (A \cap B)/(A \cup B) \quad \text{eq. (3)}$$

If user A added user B to her trusted network (i.e, A points to B), the inlink power (impact) of the user B for the user A represents how much the information of user A is influenced by the information of user B. The outlink power of User B is how much the information of user B affects the user A. The overall power measures the fraction of overlapped information in the joint information space of both users.

For the information similarity in trusted network, the following hypotheses were assessed: **H1.** Users connected by direct or indirect relationships of trust have more similar information items and meta-data than a non-connected pairs. **H2.** Users in reciprocal relations have more similar information item and meta-data than users in unidirectional relations.

5. THE RESULTS

5.1 Information sharing in trusted network

To test whether users connected by direct or distant links of trust share more information than non-connected pairs (H1), we compared both absolute numbers of shared information items and their normalized numbers (inlink, outlink, and overall powers) using one-way ANOVA test.

First, we explored the number of shared items and meta-data. Table 2 shows mean numbers of shared items and meta-data for direct and distant relationship on contrast to a non-related pair of users (which we can interpret as infinite distance). At average, direct pairs share the largest number of items and meta-data. The numbers are decreasing with the increase of distance in the network of trust achieving its minimum for non-connected pairs. This is the evidence that users connected in a network of trust do have significantly more similar interests than non-connected users. We can also consider it as an evidence of information propagation along a network of trust, although impressive similarity on the meta-data level (which are hard to propagate!) hints that interest similarity may play a more important role than propagation in the observed phenomenon. As Table 2 shows, reciprocal relationships exhibit the same pattern, also with significant differences between columns in the number of shared information items and meta-data.

Table 2. The Average Number of the Common Information

		Direct	1hop	2hops	No Rel.
Unidirect-ional	Items	**.82**	**.20**	**.14**	**.00**
		$F(3, 412315) = 6961.18, p < .001$			
	Meta-data	**22.65**	**18.85**	**20.04**	**.02**
		$F(3, 412315) = 618.37, p < .001$			
Reciprocal	Items	**8.35**	**1.50**	**.72**	
		$F(2, 1368) = 137.40, p < .001$			
	Meta-data	**93.02**	**67.77**	**33.59**	
		$F(2, 1368) = 9.16, p < .001$			

Second, we explored differences between relative similarity measures – fractions of shared items and meta-data for unidirectional relationship (Table 3) and reciprocal relationships (Table 4). In both cases, same pattern can be observed for relative similarity measures: directly related users have the largest fraction of shared items and meta-data and the fractions decrease with the increase of the distance between users and reach the minimal level for not connected users (infinite distance).

Table 3. The Average Similarity Powers of Common Information (Unidirectional Relations)

		Direct	1hop	2hop	No Rel.
Items	Inlink	**2.01%**	**0.55%**	**0.41%**	**0.03%**
		$F(3, 412315) = 2841.92, p < .001$			
	Outlink	**0.85%**	**0.17%**	**0.10%**	**0.00%**
		$F(3, 401164) = 5643.51, p < .001$			
	Overall	**0.35%**	**0.07%**	**0.04%**	**0.00%**
		$F(3, 412315) = 7696.06, p < .001$			
Meta-Data	Inlink	**5.33%**	**1.64%**	**1.54%**	**0.02%**
		$F(3, 412315) = 1969.01, p < .001$			
	Outlink	**2.87%**	**2.88%**	**2.74%**	**0.03%**
		$F(3, 401164) = 1383.66, p < .001$			
	Overall	**1.25%**	**0.77%**	**0.76%**	**0.01%**
		$F(3, 412315) = 908.51, p < .001$			

Table 4. The Average Similarity Powers of Common Information (Reciprocal Relations)

		Direct	1hop	2hop
Items	Inlink & Outlink	**6.79%**	**1.05%**	**0.37%**
		$F(2, 1368) = 160.70, p < .001$		
	Overall	**2.45%**	**0.32%**	**0.10%**
		$F(2, 1368) = 258.52, p < .001$		
Meta-data	Inlink & Outlink	**13.01%**	**6.48%**	**3.20%**
		$F(2, 1457) = 36.08, p < .001$		
	Overall	**4.79%**	**2.47%**	**1.26%**
		$F(2, 1456) = 23.92, p < .001$		

In addition to demonstrating a clear connection between item and meta-data level similarity and user closeness in a network of trust, the data shown above allows to make interesting observations. First, as we expected, between-user similarity on the level of *meta-data* is much larger than similarity on the level of *items* for both systems. For example, the inlink power similarity of items in direct relation is 2.01% while inlink power similarity of meta-data in the same direct relation was 5.33%. Second, both absolute and

relative similarities are pair-wise larger for reciprocal than for unidirectional connections for all distance levels. This difference is most pronounced in relative form reaching its highest level for direct reciprocal relations (6.79% for items and 13.01% for metadata). Next section examines the difference between reciprocal and unidirectional connections in details and checks its significance.

5.2 Unidirectional vs. Reciprocal Relations

To compare the differences of information sharing pattern between unidirectional and reciprocal relations, we started with comparing the number of shared information items and meta-data, doing it now separately for several distances of relations. In all three distances but meta-data of 2-hop connection, the numbers of shared information items and meta-data in reciprocal relations were significantly larger than in unidirectional relations. In case of meta-data of 2-hop relation, there was no significant difference.

Secondly, we checked the significance of observed differences in relative information item similarity between reciprocal and unidirectional relations (Table 6). For direct and 1-hop relationship, the differences appeared to be significant, i.e., users connected by a direct or 1-hop distanced reciprocal relation shared significantly larger fractions of information items than users connected by unidirectional relation. For 2-hops relations the observed difference appeared to be non-significant for one out of three relative similarity measures.

Table 5. Results for Powers of Information Items

		df	*t*-value	Sig.
Inlink Power	Direct	11478	-7.39*	p < .001
	1Hop	17787	-2.57*	p = .010
	2Hop	30568	.321	p = .748
Outlink Power	Direct	11478	-21.35*	p < .001
	1Hop	17787	-14.05*	p < .001
	2Hop	30568	-7.70*	p < .001
Overall Power	Direct	11478	-21.09*	p < .001
	1Hop	17787	-13.08*	p < .001
	2Hop	30568	-5.93*	p < .001

On the final step we compared relative information meta-data similarity for reciprocal and unidirectional relations (Table 7). The relative source similarity was significantly higher for users connected by direct and 1-hop reciprocal relation than for users connected by unidirectional relations of the same distance. Two out of three relative similarities were significantly larger for reciprocal relations.

Table 6. Test Results about Power of Meta-data

		df	*t*-value	Sig.
Inlink Power	Direct	11356	-6.83*	p < .001
	1Hop	17596	-10.66*	p < .001
	2Hop	30597	-4.82*	p < .001
Outlink Power	Direct	11356	-12.79*	p < .001
	1Hop	17596	-6.11*	p < .001
	2Hop	30597	-1.00	p = .318
Overall Power	Direct	11356	-9.71*	p < .001
	1Hop	17596	-7.52*	p < .001
	2Hop	30597	-2.78*	p = .005

6. CONCLUSION AND DISCUSSION

To prove the feasibility of users' self-defined relations of trust as the bases of recommendation, we examined how similar interests of users connected by a self-defined relation of trust are. Using *CiteuLike* datasets, we found that user connected by a self-defined relation of trust have more common information items and meta-data than user pairs with no connection. The similarity was largest for direct connections and decreased with the increase of distance between users in the network of trust. Users involved in a reciprocal relationship exhibited significantly larger similarity than users in a unidirectional relationship on all levels. Moreover, similarity on the level of meta-data (authors) was larger than similarity on the level of individual items (references).

While the results of our study support the idea of using networks of trust in CF systems, they still do not answer the question how to use this information to improve the quality of recommendation. In out future studies we plan to address this issue. As the first step, we will investigate the impact of trusted networks on recommendation quality using our *CiteuLike* data set. We will also explore how information propagates within trusted networks and investigate the influence of information authorities who play a leading role in disseminating the information. In later studies, we plan to expand our target domains by adding different data sets.

7. REFERENCES

[1] Adamic, L. A. & Adar, E. (2003) Friends and neighbors on the Web, *Social Networks*, 25 (3), pp. 211 ~ 230.

[2] Bonhard P. & Sasse M. A. (2006) Knowing me, Knowing you - Using Profiles and Social Networking to Improve Recommender Systems, *BT Technology Journal*, Vol. 25 (3)

[3] Lam, S. K. & Riejl, J. (2004) Shilling Recommender Systems for Fun and Profit, In: *Proceedings of World Wide Web 2004*, New York, NY, USA, pp. 393 ~ 402

[4] Massa, P. & Avesani, P. (2004) Trust-aware Collaborative Filtering for Recommender Systems, In: *Proceedings of Federated International Conference On The Move to Meaningful Internet*, Agia Napa, Cyprus, pp. 492 ~ 508

[5] Massa, P. & Avesani, P. (2007) Trust-aware Recommender System, In: *Proceedings of Recommender System 2007*, Minneapolis, MN, USA, pp. 17 ~ 24

[6] Maltz, D. & Ehrlich, K. (1995) Pointing the Way: Active Collaborative Filtering, In, Proceeding of CHI' 95, Denver, CO, USA, pp. 1 ~ 8

[7] McKnight, D. H. & Chervany, N. L. (1996). The Meanings of Trust. University of Minnesota, http://misrc.umn.edu/wpaper/WorkingPapers/9604.pdf (accessed on July, 2008)

[8] Mehta, B., Hofmann, T., Nejdl, W. (2007) Robust collaborative filtering, In: Proceedings *of Recommender System 2007*, Minneapolis, MN, USA, pp. 49 ~ 56

[9] O'Donovan, J. & Barry, S. (2005) Trust in Recommender Systems, In: *Proceedings of the 10th International Conference on Intelligent User Interfaces*, San Diego, California, USA, pp. 167 ~ 174.

[10] Schafer, J. B., Frankowski, D., Herlocker, J. & Sen, S. (2007) Collaborative Filtering Recommender System, In: Brusilovsky, P., Kobsa, A. & Nejdl, W. (Eds.) *The Adaptive Web: Methods and Strategies of Web Personalization*, pp. 291 ~ 324

[11] Singla, P. & Richardson, M. (2008) Yes, There is a Correlation - From Social Networks to Personal Behavior on the Web, In: *Proceeding of the 17th International World Wide Web Conference*, Beijing, China.

[12] Sinha, R. & Swearingen, K. (2001) Comparing Recommendations Made by Online Systems and Friends, In: *Proceedings of DELOS Workshop on Personalisation and Recommender Systems*

[13] Tintarev, N. & Masthoff, J. (2007) Effective explanations of recommendations: user-centered design. In: *Proceedings of Recommender System 2007*, Minneapolis, MN, USA, pp. 153-156

[14] Ziegler, C. & Golbeck, J. (2007) Investigating interactions of trust and interest similarity, *Decision Support Systems*, 43, pp. 460 ~ 475