

How to Measure the Information Similarity in Unilateral Relations: The Case Study of *Delicious*

Danielle Lee
School of Information Sciences,
University of Pittsburgh
135 N. Bellefield Ave., Pittsburgh, PA 15260
hyl12@pitt.edu

ABSTRACT

In this paper, I describe a better way to compute the information similarity between two users who are unilaterally connected. Unilateral relations are unidirectional connections and gain attention with the success of social tagging and microblogging systems. The relations are convenient and less bounded since people can make the connection without mutual agreement once they perceive that other users' information is worth. Using a social bookmarking data set, *Delicious*, I found that the traditional item unit-based similarity measures are not enough to show the common interests between a pair of unilaterally connected users. The similarity measure on the higher level such as metadata (root address of each URL) and macro-level tags (tags regardless of the annotated information item) showed better results. The significantly better results in metadata and macro-tag level similarity were also shown in the indirect relations, as well. I interpreted this result to mean that semantic information such as metadata and tags represent users' cognitive understanding of corresponding information. Therefore, in social tagging systems, it is better to match users not based on item-level similarity but based on the similarity on a higher level which embeds more semantic meanings.

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

Social Networks, Information Sharing, Similarity Measure, *Delicious*

1. INTRODUCTION

The power of collaborative filtering (CF) recommendations is based on a relatively simple idea: target users will like items favored by their likely-minded peer cohorts. While generations of CF systems have proven the effectiveness of this approach, they provided no ways for the user to control the recommendation mechanism. Their target users are not involved in the process of

choosing the peer group and further generating recommendations. Unlike early-day "push" and "pull" collaborative filtering systems [9-10] which connected people openly, modern automated collaborative filtering systems do not even expose information about the people whose ranking was used to generate the recommendations. This lack of control over the selection of peers has disadvantages from both sides. Target users can neither add specific known and trustworthy persons to the set of their peers, nor can they exclude some strange or suspicious "peers" from the set picked up by the system. The latter inability leaves users defenseless against various spamming or "shilling" behaviors by the CF recommenders. If the anonymous group of people tends to make a profit or distort the system by using any of several known approaches [6], the end users cannot protect themselves. As a way to give users to control the recommendation process, I suggest to fuse users' own social networks with the traditional CF technology.

Compared with the era when computer users stayed in isolation, users on the Web 2.0 have found it easier to know who knows what through engaging in social networks in virtual space [12]. Unlike social networking systems (SNS) mainly focusing on linking mutually agreed-up friendships, many social tagging systems, which aim to manage interesting information online, offer a different kind of sociability – unilateral relations. These relatively new social relations gained attention along with the success of social tagging and micro-blogging applications [2-3, 5]. The most typical examples of unilateral relations are "following" on *Twitter*, "watching" on *Citeulike*, or "contacts" on *Flickr*. These relationships are solely based on users' own perceived usefulness or interests in the information. Once a user perceives that the information of another user is useful or interesting, the user is able to watch him as his follower. Then the information of the followed person is automatically displayed the followers. The relations don't require any offline interactions or emotional bonds to make connections or the mutual agreement for being connected. Wellman suggested that, in Web 2.0 era, various new relationships would emerge and the networks are "less bounded [12]." The unilateral relationship is one kind of the new relationships. Unlike befriending in SNS (e.g. facebook, MySpace or Friendster) where users increase the number of friends simply for fun or curiosity [4], the unilateral relationships aim to acquire information from the connected people's collection. When a user unilaterally follows many users, it may cause rapid growth and potential dilution of list of the items in the follower's collections. Therefore, unilateral relationships require users to be careful to select people to follow, based on the utility of information.

In this paper, I explore how social tagging systems can measure the similarity between two users in a unilateral network in a better

Permission to make digital or hard copies of all or part 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. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MSM'10, June 13, 2010, Toronto (ON, Canada)

Copyright © 2010 ACM 978-1-4503-0229-6/10/06 ...\$10.00.

way and give a hint about how to exploit these convenient social networks in personalizing information. If a user finds that another user's information collection is interesting and wants to see the updated list, we may be able to infer the following users' interests by taking into consideration the information of the followed user. Herein, I am trying to find good ways to interpret the implicit user's preferences using unilateral relations.

The CF technology may calculate the similarity in a different way than how users process information. The CF recommendations compute the item-level similarity. Once two users have exactly the same item or give the same ratings to one item, it is assumed that they are similar and their information is useful to each other. However, the item unit-based similarity may ignore the fact that users can perceive the same item with dissimilar meaning or the different items with similar meaning. In this paper, I will examine whether item unit-based similarity is the good way to measure the actual overlap of interests between two users in a unilateral relationship or not.

In addition, many social tagging systems don't require the users to rate items. While typical recommendation algorithms choose the most appropriate items based on the user's ratings given to the items, without any rating mechanism, the recommendations for social tagging systems only rely on unary ratings (i.e. whether it is tagged). Unary ratings are too simple to infer each user's preference and further compute the similarity of interests between two users. Even though it would be possible to infer binary ratings (i.e. positiveness or negativeness) through sentiment analysis of the assigned tags, users usually do not annotate anything to items that they don't like. That is to say, most of the items may be counted as positive. That is the reason why we need more semantic and multi-dimensional approaches to calculate the information similarity. In social tagging systems, users add annotations to items using their own vocabulary according to how they cognitively comprehend the corresponding item [1].

Breslin and Decker (2007) said that the social networks connecting via items of interests, which are called object-centered sociality, may be more long-lived relations than the relationships not sharing any item of interest. In order to help users develop better relationships, the authors insisted that SNS have to take into consideration people's actions about content such as tagging, blogging, adding comments, etc. to find out the users' items of interests [4]. Bojars and colleagues (2008) also emphasized the semantic aspects of social networks. They proposed a semantic Web framework (Semantically Interlinked Online Communities, SIOC) to make varied semantic information from miscellaneous sources interchangeable. The authors suggested that their own framework could be used together with Friend-of-a-Friend (FOAF) to show social connections. However, their approach is based on social connections via shared objects only, and they did not show any exemplary system which applied both SIOC and FOAF [11]. Even though many social tagging systems support the similar kind of unilateral or following relations, surprisingly, there are few studies about how to utilize this social network for personalizing information. Java and the others (2007) insisted that one of the reasons why users enjoy microblogging, such as on the Twitter, is to share information. They also found three kinds of users' intention to socialize on the Twitter (i.e. *following*); not only friendship-wise relationships but also information sharing and information seeking [3].

2. THE DATA SET

2.1 The Data Source and the Relationships

As a source of data for my study I selected a collaborative tagging system *Delicious*¹. Like many other collaborative tagging systems, *Delicious* supports unilateral relationships, and the relationships are called "network" in *Delicious*. When user A considers a collection of bookmarks which are assembled by another user, user B as a good source of information, he can *network* with that user without user B's consent. Then, all items added to user B's collection are automatically displayed to user A.

The networking relationship was the primary focus of investigation in my study. In my analysis, I collected directly and indirectly networking relations. The direct relations consist of two nodes which are linked without any intermediate path as shown on Figure 1. The *networking user* is connected to the *networked user*, but the relationship is not necessarily reciprocal. For instance, user A is referring to user B via *network*. Then, user A is the *networking user* and user B is the *networked user*. The indirect relations are the people who are associated with one another through paths between them as shown on Figure 2. In this study I explore this indirect relation so as to investigate the transitivity of user's interests. I collected relations with two distances between users: one hop and two hops (Figure 2).



Figure 1. Direct Relation in Networking Relationship

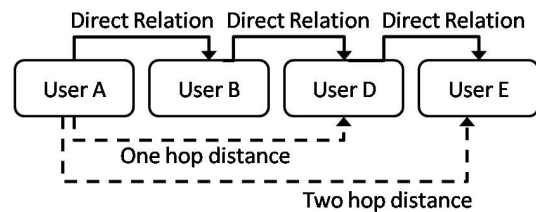


Figure 2. Indirect Relation in the Center of User A

In order to collect the data set, I used snowball sampling. To pick the initial seed set of users, I visited *Delicious* randomly on the first and second week of November 2007. All users who posted new bookmarks at the time of visit were chosen. For each user, I collected bookmarks, the tags, and user *networks*. After selecting a group of initial users, I performed the breadth-first search to collect their networking users and the data. Table 1 is the descriptive statistics of the data set. Figure 3 shows the number of networked users to whom each user was connected. Many users were connecting to only a small number of users, but some users connected to a lot of users (four users who were networking to more than 100 users) with some exceptions.

2.2 Measures in Consideration

As the measures to calculate the information similarity between two users in unilateral relations, I took into account not only the information item-level similarity as a control variable, but also metadata-level and tags-level similarity. Herein, metadata represents origins of information items. Specifically, I considered

¹ <http://www.delicious.com/>

Table 1. Data Summary of Dataset

Total no. of users	11772
Total no. of distinct items (bookmarks)	5191538
Average no. of items per user	788.83
Total no. of distinct tags	651622
Average no. of tags per user	2268.87
Total no. of unilateral relations	16538

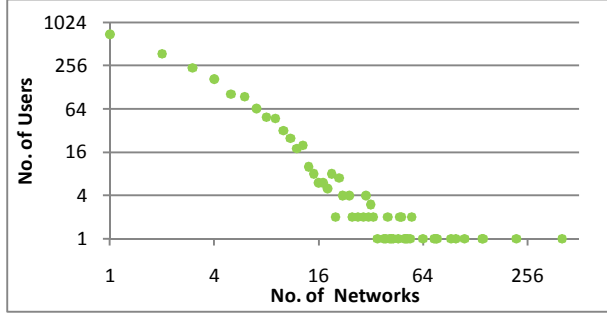


Figure 3. Distribution of Unilateral Relations

root address of each bookmark (i.e., Web site it came from) as its origin. I assume that information items coming from the same origin could be similar semantically or contain similar characteristics. I hypothesized that the users who share the same interests may not necessarily agree on specific shared items, but demonstrate higher agreement on the level of information metadata. In tag-based similarity, I considered two aspects: *micro* level and *macro* level tag similarity. Micro level represents tag similarity for a common information item and macro level represents the similarity of the overall collections of tags of two users, regardless of the tagged information item.

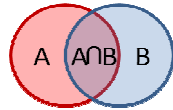


Figure 4. Information Overlap

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

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

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

Since the size of users' items and tags collection vary from user to user, I counted both the absolute and relative measures on all the above four levels (i.e. item, metadata, micro-tags and macro-tags level). For the absolute measures, I counted the raw number of common information. For the relative measures, I counted inlink power, outlink power, and the Jaccard similarity. The former two variables measure the direction of influence. For example, user A is networking to user B, and user A and user B have 50 items and 100 items, respectively. If there are 10 common items, 20% of user A's collection is overlapped with user B's collection and 10% of user B's collection is overlapped with user A's collection. Depending on the direction of the linkage, the similarities are different. The *inlink power* (eq. 1) is the portion of shared information on the center of a networking user. The *outlink power* (eq. 2) is the portion of shared information on the center of a networked user. I assumed that networking with another user may be more meaningful to the networking users than the networked users. It is because the watching users feel interest in other watched users' collection, but not vice versa. In a later section, I

will examine whether these two variables were really different meaning to users. The Jaccard similarity is the fraction of shared information on the center of the joint information space of both users (eq. 3). In order to test the significance of the mean difference among the measures in the test and to find out the best measure, I used Friedman two-way ANOVA procedure with a significant level of 0.05.

3. THE RESULTS OF DATA ANALYSIS

In the following section, I tested which measure is a better one to represent the similarity of interests between two people in a unilateral network. As the measures, I explored the similarities of absolute numbers and relative numbers in four explored levels (item, metadata, micro-tag and macro-tag). Before starting the analysis of the similarities, I examined general patterns of collecting information. I compared the size of the collections between networking users and networked users. The networking users ($M = 1520.4, \sigma = 1491.1$) owned larger collections than the networked users ($M = 1123.5, \sigma = 1300.6$). This mean difference was statistically significant according to the results from the Mann-Whitney nonparametric test (Mann-Whitney $U = -25.43, p < .001$). Since the standard deviation is large, I also counted the differences in the size of collections that every pair of users have. Out of 16538 pairs, networking users of 9719 pairs (58.8%) are richer than their networked users and networking users of 6819 pairs (41.2%) are poorer than their networked users. Therefore, the tendency that the networking users are richer than the networked users is not obvious.

As the first test of similarity measures, I compared the similarity based on absolute number (refer to Table 2). As the results show, the similarity on the macro-tag level was the largest, and the one on the metadata level was the second largest. However, the item based similarity was relatively trivial compared with the metadata and macro-tag similarity. The differences among the means were statistically significant ($\chi^2 = 37523.68, p < .001$).

Table 2. Similarities of the Absolute Numbers

Item	Metadata	Micro-tag	Macro-tag
12.83	50.14	12.65	97.83
($\sigma = 29.51$)	($\sigma = 69.05$)	($\sigma = 41.50$)	($\sigma = 127.23$)

Although the raw number of common macro-tag and metadata was quite large, as aforementioned, the number of items each user has vary user by user. Therefore, I calculated relative similarities. As Table 3 shows, the results of all relative similarity measures were consistent with the results of absolute numbers. The macro-level tags and metadata represented the larger similarity than the item and micro-level tags.

I interpret this result to mean that users perceived the information utility of another user's collections from higher level such as the background semantics than individual item level. If the semantic information is important, which relative measure is the most meaningful to represent the similarity? I compared three relative measures, Jaccard similarity, inlink and outlink power, on macro-tag level and on metadata level. The outlink power on macro-tag level was significantly larger than the inlink power and the Jaccard similarity ($\chi^2 = 24396.3, p < .001$). As the next step, I counted the number of pairs whose outlink power on macro-tag is larger than the inlink power. 57.2% of pairs have the larger outlink powers and 42.8% of pairs have the larger inlink powers. The result was the same with the differences in the size of collections between the networking users and the networked users

explained above. When I computed the metadata-level similarity, the outlink power was significantly larger than the inlink power and the jaccard similarity, as well. That is to say, in order to compute the information similarity of unilateral connections, it is important to consider not only the similarity centered on the networking users' collection (inlink power), but also the similarity centered on the networked users' collection (outlink power).

Table 3. Similarities of the Relative Numbers

	Item	Metadata	Micro-tag	Macro-tag
Jaccard Similarity	0.42% ($\sigma=0.84\%$)	2.7% ($\sigma=2.5\%$)	0.13% ($\sigma=0.4\%$)	7.31% ($\sigma=5.0\%$)
Inlink Power	1.47% ($\sigma=4.60\%$)	7.94% ($\sigma=9.35\%$)	0.50% ($\sigma=2.41\%$)	18.62% ($\sigma=16.13\%$)
Outlink Power	1.99% ($\sigma=5.47\%$)	10.62% ($\sigma=11.22\%$)	0.68% ($\sigma=2.79\%$)	23.58% ($\sigma=18.21\%$)

Table 4. Similarity Measures for One-hop Indirect Relations (with Std. Deviation and the Statistical Test Results)

	Item	Metadata	Micro-tag	Macro-tag
Absolute Numbers	7.40 ($\sigma=17.8$)	50.50 ($\sigma=64.6$)	6.10 ($\sigma=19.8$)	109.15 ($\sigma=124.2$)
	$\chi^2 = 25733.3, p < .001$			
Jaccard Similarity	0.17% ($\sigma=0.4\%$)	2.19% ($\sigma=1.9\%$)	0.14% ($\sigma=0.7\%$)	7.0% ($\sigma=4.2\%$)
	$\chi^2 = 26671.0, p < .001$			
Inlink Power	0.50% ($\sigma=1.2\%$)	6.58% ($\sigma=6.6\%$)	0.14% ($\sigma=0.7\%$)	18.59% ($\sigma=14.5\%$)
	$\chi^2 = 27386.2, p < .001$			
Outlink Power	0.61% ($\sigma=1.2\%$)	7.71% ($\sigma=7.4\%$)	0.17% ($\sigma=0.5\%$)	20.05% ($\sigma=15.0\%$)
	$\chi^2 = 27279.7, p < .001$			

Table 5. Similarity Measures for Two-hops Indirect Relations (with Std. Deviation and the Statistical Test Results)

	Item	Metadata	Micro-tag	Macro-tag
Absolute Numbers	6.08 ($\sigma=15.7$)	43.97 ($\sigma=59.1$)	4.59 ($\sigma=15.1$)	101.08 ($\sigma=115.9$)
	$\chi^2 = 57652.9, p < .001$			
Jaccard Similarity	0.14% ($\sigma=0.3\%$)	1.90% ($\sigma=1.6\%$)	0.04% ($\sigma=0.1\%$)	6.69% ($\sigma=4.0\%$)
	$\chi^2 = 61240.4, p < .001$			
Inlink Power	0.40% ($\sigma=0.9\%$)	5.51% ($\sigma=5.5\%$)	0.10% ($\sigma=0.3\%$)	17.45% ($\sigma=13.6\%$)
	$\chi^2 = 60730.5, p < .001$			
Outlink Power	0.54% ($\sigma=1.2\%$)	7.04% ($\sigma=6.8\%$)	0.15% ($\sigma=0.4\%$)	19.34% ($\sigma=14.8\%$)
	$\chi^2 = 60304.9, p < .001$			

I tested whether the macro-tag based and metadata based relative similarities can be applied to the indirect relations and have transitive powers or not. The Table 4 shows the results of absolute measure and relative measures for one hop distance relations, and Table 5 shows the results for the relations with two hops. Even though, as the network distance increased, the similarity power became smaller, the results showed the consistent patterns with the result of direction relations. For both of indirect relations, the macro-tag level similarity was the best and the metadata level similarity was the second best measure. These were statistically significant. Said differently, the relative measures on macro-tag level and metadata level were more important than item-unit level similarity also for the relations with distances. When I counted the ratio of pairs whose outlink powers were larger than the inlink

powers, it was 52.2% of one-hop relations and 52.4% of two-hop distance relations. Therefore, like the direct relations, it is better to take into account both inlink and outlink powers for distanced relations. Lee & Brusilovsky found that a pair of users who are in unilateral relations shared more common information than another pair of users who are not connected at all [7-8].

4. CONCLUSION AND DISCUSSION

In this study, I compared various measures to compute the similarity between two nodes in unilateral relations. The test results of direct relations showed that higher-level information similarity such as macro-tags and metadata were more important measure to compute the similarity than item-unit based similarity. This pattern was consistently shown in the results for the indirect relations. Therefore, when calculating the interest similarity between two users in a unilateral relationship, semantically rich information is important.

As the future direction, it is necessary to compare the information similarity between unilateral relations and peer cohorts, which are chosen by item-unit based automated CF approach. I will develop recommendation algorithms based my findings in this paper, as well. In addition, in order to reinforce the result of this study, I plan to add different social tagging data sets.

5. REFERENCES

- [1] Golder, S.A. and B.A. Huberman, *Usage patterns of collaborative tagging systems*. J. Inf. Sci., 2006. **32**(2): p. 198-208.
- [2] Huberman, B.A., D.M. Romero, and F. Wu, *Social networks that matter: Twitter under the microscope*. First Monday, 2008. **14**(1-5).
- [3] Java, A., et al. *Why we twitter: understanding microblogging usage and communities*. in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*. 2007. San Jose, California: ACM.
- [4] John, B. and D. Stefan, *The Future of Social Networks on the Internet: The Need for Semantics*, in *IEEE Internet Computing*. 2007. p. 86-90.
- [5] Johnson, S., *How Twitter Will Change the Way We Live*, in *Time*. 2009, Time.
- [6] Lam, S.K. and J. Riedl, *Shilling recommender systems for fun and profit*, in *Proceedings of the 13th international conference on World Wide Web*. 2004, ACM: New York, NY, USA. p. 393-402.
- [7] Lee, D.H. and P. Brusilovsky. *Does Trust Influence Information similarity?* in *Proceedings of the ACM RecSys'09 Workshop on Recommender Systems & the Social Web*. 2009. New York, NY, USA.
- [8] Lee, D.H. and P. Brusilovsky. *Social Networks and Interest Similarity: The Case of CiteULike*. in *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*. 2010. Toronto, Canada.
- [9] Maltz, D. and K. Ehrlich, *Pointing the way: active collaborative filtering*, in *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1995, ACM Press/Addison-Wesley Publishing Co.: Denver, Colorado, United States. p. 202-209.
- [10] Schaefer, J.B., et al., *Collaborative Filtering Recommender Systems*, in *The Adaptive Web: Methods and Strategies of Web Personalization*, P. Brusilovsky, A. Kobsa, and W. Nejdl, Editors. 2007, Springer: Berlin, Germany. p. 291-324.
- [11] Uldis, B., et al., *Interlinking the Social Web with Semantics*. 2008. p. 29-40.
- [12] Wellman, B., *Computer networks as social networks*. Science, 2001. **293**(5537): p. 2031-4.