How to Measure Information Similarity of Online Social Networks:

A Case Study of Citeulike

Danielle H. Lee, PhD*

Adaptive Interactions, Co.

1460-1 Bangbae-dong, Seocho-gu, Seoul, South Korea, 06568

suleehs@gmail.com

Peter Brusilovsky, PhD

School of Information Sciences, University of Pittsburgh

135 N. Bellefield Ave., Pittsburgh PA., USA, 15260

peterb@pitt.edu

*** indicates corresponding author.**

ABSTRACT

In our presently knowledge-driven society, many information systems encourage users to actively utilize their online social connections' information collections as useful sources. The abundant information-sharing activities among online social connections could be valuable in enhancing and developing a sophisticated user information model. In order to leverage the shared information as a user information model, our preliminary job is to determine how to effectively measure the resulting patterns. However, this task is not easy due to multiple aspects of information and the diversity of information preferences among social connections. Which similarity measure is the most representable for the common interests of multifaceted information among online social connections? This is the main question we will explore in this paper. In order to answer this question, we considered users' self-defined online social connections, specifically in Citeulike, which were built around an object-centered sociality, as the gold standard of shared interests among online social connections. Then, we computed the effectiveness of various similarity measures in their capabilities to estimate the shared interests. The results demonstrate that, instead of focusing on monotonous bookmark-based similarities, it is significantly better to zero in on more cognitively expressible metadata-based similarities in accounting for the shared interests.

Keywords

Information Similarity, Online Social Networks, Watching Relations, Group Membership, Social Tags, Citeulike

1. INTRODUCTION

Over the last 10 years, online social networks have evolved into a major medium to help users not just to socialize but also to find useful information. The pervasiveness of online social networks leads many information systems to adopt *online sociability* as assistance for users to find and share useful information. The features of modern video- or photo-sharing sites (e.g., Youtube subscription; Flickr contacts and groups; Instagram's following), social bookmarking systems (e.g., Citeulike watching and groups; Delicious network; Mendeley groups) and blogging or micro-blogging sites (i.e. Twitter following; Naver neighbors) are good examples of this trend. In other words, contemporary information systems deliberately encourage their users to connect to like-minded peers and utilize information flows or information collections of peers as valuable information sources. At the same time, users of social systems are quite motivated to connect with similar users who might serve as sources of valuable information [5].

In this context, the ability of supporting users to recognize his or her like-minded peers (who, consequently, can serve as sources of valuable information) emerges as an important functionality of social systems. The functionality is critical for several advanced information access approaches such as adaptive navigation support, adaptive search, or personalized recommendation [50]. For example, once peer users are identified, the social system can directly examine the collections of

these peers and recommend the most promising items from these collections to the target user. The approach echoes the popular collaborative filtering technology where users with similar ratings are considered as peer users, and their positively rated items are used for recommendation [50].

The main problem is how to determine truly similar users in social media systems where users create a variety of information types. Classic collaborative filtering recommender systems rely on item ratings submitted by user and use various rating similarity measures to identify like-minded peers. However, most of social media systems provide no rating functionality. The users in social media systems usually share pieces of information (e.g., add tweets, post photos and messages, add new research papers, or share videos) and *adopt* pieces shared by others (by copying, liking, re-tweeting, etc) without any explicitly defined degree of preference. To find true peers in the absence of reliable ratings, social systems need to consider various aspects of information-sharing patterns among users to derive an ideal similarity measure of common interests. The problem here is that there are multiple information sources and parameters that could be potentially useful to estimate common interests of social system users based on the patterns of sharing non-rating contents. Should we ignore item content (treat an item as a black box) and only consider the number of commonly shared items between users like other existing studies? Otherwise should we take into account the size of a user library when deducing similarity from shared items? Alternatively, should a similarity measure go beyond overlap measures and take into account the *content* and *metadata* of shared items? For instance, when recommending academic articles, titles, abstracts, keywords, authors, and published journals/conferences are important characteristics [34]. Thus, the presence of the same authors or keywords in collections of two users could indeed indicate the overlapped research interests between the users.

The idea to focus on the number of common items between users without any consideration of content and metadata is the most close analogy to the current rating-based approach in collaborative filtering and was first to be explored [47]. However, it has been also argued that content-oriented similarity approaches are more meaningful for social systems where the volume of socially shared content (e.g., tweets, photos or bookmarks) is much larger than the volume of items (e.g., movies or books) in traditional recommender systems [41]. As a result of this large information flow, like-minded users tend to have a much lower chance to share or adopt exactly same items because it is infeasible to look through all available information [13]. Although a number of similarity approaches of both kinds (i.e., based on the count of common items and based on the content or metadata of common items) were suggested in the past [36], no systematic evaluation and comparison of these approaches has been performed. To bridge this gap, this paper aims to comparatively analyze a variety of similarity measures in a social system with diverse information to determine an approach that offers the best chance to identify truly similar users.

The main challenge of such a comparative analysis is to identify a gold standard of similarity (i.e., reliable knowledge about true user similarity) that could be used to evaluate the quality of various similarity approaches based on sharing patterns. This is where object-centered social networking functionality provided by many social sharing systems can help. Indeed, this functionality enables users to establish connections with other users for the purpose of information sharing [9]. Good examples of object-centered social networking are the ability to unilaterally "follow" other users (i.e., 'following' on *Twitter*, 'network' on *Delicious,* 'circles on *Google Plus)* or join online groups. These connections are self-established and predominantly content-oriented, because they support content access but no personal and social interactions [9]. In this situation, the self-established social connections could serve as a gold standard for evaluating similarity measures because they are likely to link the target user to truly similar likeminded users.

In our analysis of user similarity measures in social systems, we use data from the social bookmarking system *Citeulike*, which enables users to establish two types of information-oriented social connections: unilateral following (called '*watching*' in *Citeulike*) and group co-membership. To make a more reliable comparison of various similarity measures, we use each of these self-established connections as a gold standard (Section 5). We start our work with examining the properties of these connections to confirm that they could, indeed, serve as gold standards. Once it is confirmed, we proceed to compare similarity measures by examining which measure can most effectively approximate the gold standard similarity (Section 6). To empirically test the effectiveness, we generated recommendations of social connections using each similarity measure and counted which measure best predicts the existing gold standard social connections. We complete this paper with a discussion and conclusion (Section 7). Taken together, the contribution of our paper is an examination of candidate gold standards for measuring user similarity in social systems and a systematic comparative evaluation of several popular similarity measures in order to find the optimal similarity measure that could be used to identify genuinely like-minded users in social systems.

2. RELATED WORKS

2.1 Online Social Networks and Interest Similarity

Table 1 summarizes existing studies that investigated the information similarities of online social networks. As shown, these studies investigated how online social connections share similar information from a limited perspective – mostly based on the count of commonly bookmarked or rated items – with little consideration that a user's information preferences could be exposed in a variety of ways. In addition, most of them focused on just one type of online social connections – online friendships – and neglected the diversity of social connection types available in modern social systems.

Table 1. Studies about Information Similarity of Online Social Networks

Paper	Input of Similarity Measures	Similarity Measures	Kinds of Social Networks	Domain
Akcora & Carminati [1]	Friendship Graph	Social Structure-based Similarity	Friendship	Online SNS (Facebook)
Anderson, et al. [2]	Users' actions (i.e. editing posts, questioning and answering and evaluating reviews)	Co-occurrence of same user actions	Implicit Online Social Connections	Wikipedia, Stack Overflow & Epinions
Baartarjav, et al. [4]	Personal traits such as age, gender, religion living are, political opinions, etc.	Clustering of personal traits	Friendship	Online SNS (Facebook)
Bhattacharyya, et al. [6]	Keywords in Facebook user profile	Distance between keywords based on the 'Forest model'	Friendship	Online SNS (Facebook)
Bischoff [7]	Music Listening History (loved and banned tracks) and the Related tags	Similarity of Music Listening History	Friendship	Music (Last.fm)
Brzozowski, et al. [10]	Voting Patterns of Political Resolves	Similarity of Voting Values (i.e. ratings)	Friendships, Ideological Allies and Foes	Online Forum about Political Problems
Hajian & White [14]	Users' various activities (posting, liking and commenting) & following network	The frequency of the same activities between users and the similarity of social structure	Following Network	Online SNS (FriendFeed)
Lee & Brusilovsky [21]	Bookmarks of Scientific Articles and the Social Tags	Similarity of Commonly Bookmarked Items	Watching Relations	Social Bookmarking System (Citeulike)
Liu, et al. [24]	Ratings of Various Products	Similarity of Ratings	Trust-based Network	Product Review (Epinions.com)
Ma, H. [26]	Movie Ratings and Check-in Records of several Venues	Similarity of Ratings	Friendship and Trust-based Network	Movies & Venues
Modani, et al. [29]	5-Scale Ratings of Movies and Friendship Structure	Item Rating-based Similarity and social structure-based similarity	Friendship	Movie Review (Filmtipset)
Yu, et al. [48]	Users' following network	Social Structure-based Similarity	Following Networks	Micro-blogging Site (Weibo)

For instance, Bhattacharyya, et al. (2011) suggested a way to measure the similarity of keywords appearing on Facebook user profiles among Facebook friends. Although a variety of menus exist on a Facebook profile, and diverse personal traits are inferable from online interactions, the authors of the paper relied on a limited aspect, "interests" menu. The results of their analysis did not match the well-known definition of homophily, either. The farther two users are distant in social structure, the more they have similar keywords [6]. Because several studies already approved the homophily in Facebook [25], the effectiveness of keyword similarity based on lexical hierarchy as a measure of homophily is in question. In another study based on Facebook, Akcora and Carminati (2011) examined network-only graph-based similarities, thus focusing on how much their proposed similarity measures would predict existing social connections [1].

As an effort to add more dimensions to rating-based similarity measure (i.e. Pearson correlation coefficient, PCC hereafter), Liu, et al (2014) proposed to consider three perspectives – proximity, impact and popularity – additionally in PCC. Proximity penalizes ratings in disagreement, impact reinforces the positive and negative similarity of two given users, and popularity

denotes how common two users' ratings have. The results of personalized recommendations based on the proposed similarity measures yielded better accuracy than other established similarity measures [24]. However, our current study is based on social bookmarks where numeric ratings are unavailable.

2.2 Homophily and Social Influence

Many social scientists have suggested that we feel attractions to other people who are similar to us; thus, we selectively make social connections with them due to the ease of communication, shared knowledge, and other factors making the interactions comfortable [25]. The principle articulating this social selection made by a person's perceived similarity is referred to as *homophily* [46]. Cumulative studies about homophily have demonstrated that the similarity has been traditionally associated with personal status: for instance, age, sex, religion, ethnicity, educational and occupational class, social positions, etc. [28, p.416]. However, the contemporary society driven by information and knowledge has led into new dimensions of homophily. We feel attracted to people whose information or knowledge is useful (i.e. high utility) and whose information preferences are similar to our own. It is because people tend to perceive high intellectual values from those who are similar [16]. We use others as a reference group and compare ourselves with the reference to obtain information or make a decision as a *social comparison* process [46]. In one exemplary study, Singla and Richardson (2008) tested the relationship between instant messenger logs and the similarity of search queries. The authors were able to demonstrate that search queries of people who exchanged instant messages frequently shared more similar interests than those of random pairs. Moreover, the longer the people talked, the more similar were their queries [39].

In addition, another preeminent social science theory, *social influence*, suggests that social connections with similar people affect our ways of thinking, attitude, the information and interactions for us to choose, and other various decision-making tasks. While homophily explains a mechanism of *why* people selectively choose their social partners, the social influence explains *how* people are affected by their social partners. Social influences can be distinguished into two distinct processes according to the expected results of the influences – *normative influence* and *informational influence*. While *normative influence* is an influence to conform to the positive expectation of social connections due to a desire to be a member of a small social network, the *information influence* is to "accept information from another as trustworthy evidence about objective reality," when people are not sure about an accurate view of reality [20].

Last, in contemporary society, which is highly driven by information and knowledge, the theory of *object-centered sociality* insists that knowledge cultures are inter-stitched with current social structures. The main idea of the theory is that information-driven culture spills and weaves the fabrics of modern society, and information objects are social interaction

triggers and anchors of communications [32]. The two types of social networks that we consider in our study – watching relations and group co-memberships – have high degree of object-centered sociality [31].

3. ONLINE SOCIAL NETWORKS AND THE DATA SOURCE

This paper aims to determine the most effective way to determine users' information similarities in social sharing systems. Specifically, we compare the effectiveness of various similarity measures using two types of self-established information-centered social connections as gold standards of information similarity. This section explains the characteristics of the two online social connections and the data sets.

3.1 The Focus: Watching Networks and Group Membership

In this study, we consider two kinds of self-declared social connections existing on Citeulike – watching connections and group membership – as the gold standards to compare the effectiveness of various user similarity measures. Watching connections forming *the watching network* is a typical example of newly emerging and less-bounded online social connections. Users on Web 2.0 have found it easier to know "who knows what" through online social networks. It is a burden, however, for users to directly contact the person who has the desired knowledge via their personal ties [31]. A one-way self-established watching relationship relieves the burden of contacting other users directly for information while simplifying access to information shared by these users. Once users find other users whose interests and preferences are useful or similar, they are allowed to continuously *watch* information collections of interesting users without consent of the watched parties. Compared with SNSs (i.e., Facebook, LinkedIn, or Friendster), which focus on *mutually agreed-up* friendships, information systems offering watching networks (i.e., social tagging systems like Citeulike and Delicious and micro-blogging systems like Twitter [33]) predominantly aim to manage and share interesting information. This type of social connections is convenient and based on the utility of information possessed by users.

Because the watching connections are built around users' perceived utility of their social partners' information, we suggest that this is a highly object-centered sociality. It has been demonstrated that social networks established around items of interest are more long-lived than the relationships not sharing any item of interest [31]. Weng and the others (2013) found several kinds of user strategies for socializing on Twitter (i.e. *following*): 1) information-oriented; 2) friendship; 3) casual friendship; 4) random connections; 5) mixture of miscellaneous connecting behaviors. The authors also found that the users using information-oriented socialization strategy is inclined to stay active longer and be more active in producing and propagating contents than other users the other strategies [43]. On Citeulike, once a user starts to watch another user, the aggregated list of items collected by the user's all-watched users is automatically updated to the watching user on his *watchlist* page. This functionality enables users to declare users of interest leaving it to the system to assemble the watched

users' information in a separate page for the future reference. Hence, the watched users' information could be used as a part of the watching users' information sources.

Group co-membership is another type of self-established information-oriented social connection. Users join *online groups* to be a part of a community of interest or practice (for instance, a fan club of a musician, a community of Hadoop programmers, an online forum for students taking the same class, an online space for members of the same project, etc.). Users belonging to the same group typically distribute and contribute topic-relevant information. The theory of communal sharing relationships introduces the social dynamics of online group memberships. Group members believe that they share common substances and treat information objects as the shared substances [12]. Group members are willing to share what they need and contribute what they can. Members usually do not expect to receive something in return for their contributions. Simply being a member of a group is sufficient to them because they are able to use the resources the group is sharing [35]. Although we defined that group memberships are self-organized by the members, most studies regarding online group dynamics concentrated on the techniques to derive implicit communities using various machine-learning technologies and the characteristics of the derived implicit communities [42]. In contrast with these studies focused on implicit communities, our study seeks to explore information-sharing patterns of users' self-defined (explicit) online groups.

3.2 Data Source and the Descriptive Statistics

The data source of our study is Citeulike (http://citeulike.org), which is one of the leading systems for managing and sharing bibliographic information. This system was chosen for several reasons. First, it focuses on one kind of information item – bibliographic information of scholarly references (i.e. articles and books). Second, it provides both watching network and group memberships to help its users acquire useful information. The core component of our data was officially distributed by the Citeulike administrators, specifically, the version downloaded on May 15, 2011[1]. The data set contains all article IDs, all users, users' bookmarks, dates and times of the bookmarks and social tags at the time when this data set was made. In addition, this data set contains a list of whole groups and all members of each group. Because it does not provide metadata of articles (i.e. titles, authors, publication journals, abstracts, and publication years) and watching relations, we collected the information separately. Table 2 shows the descriptive statistics of the data set. Because our study aims to compare information similarity among users, we excluded users who have no bookmarks. Moreover, our study aims to compute information similarities among co-members of the same group. Hence, we also excluded groups having a single member.

Table 2. Descriptive Statistics of the Dataset

Watching Relations	No. of Users in Watching Relations	11,439

[1] http://www.citeulike.org/faq/data.adp

	No. of Relations	44,847
	Avg. No. of Watched Partners per Watching User	13.91 (σ = 28.7)
Groups & the Memberships	No. of Groups	1,870
	No. of Users having group membership	8,009
	No. of Group Membership	11,863
	Avg. No. of Memberships per Group	6.34 (σ = 16.3)
	Avg. No. of Memberships per Member	1.48 (σ = 1.7)
Information Collections	No. of Distinct Articles	3,210,960
	No. of Users	94,388
	No. of Bookmarks	3,869,993
	Avg. No. of Bookmarks per User	41.0 (σ = 407.5)
	Avg. No. of Tags per User	143.21 (σ =2408.3)

4. User Similarity Measures

The main purpose of this paper is to find the similarity measure that provides the best estimation of shared interests among users in social sharing systems. In search for the best measure, we consider various similarity measures based on both shared items and their metadata. In the Citeulike data set, users' preferences about information items are mainly represented by their bookmark collections. The presence of bookmarks expresses users' interests on a corresponding item, while the absence of the bookmarking does not necessarily represent that users are not interested in or disliked that item. Moreover, there is no rating or level to express the degree of users' preference. In this situation, the similarity between a pair of users can start from counting how many items they co-bookmarked. In addition, the target items of the Citeulike data are scientific articles that contain various publication-related metadata from titles and abstracts to publication years and author names. When users bookmark articles, the system also allows them to add free-text tags. Metadata and assigned tags provide an alternative way to assess item similarity [22]. Therefore, in this paper, among a variety of information similarity measures, we explore similarity measures, which are widely used in personalized recommendations and information retrieval [36, 45] to work with social bookmarks, social tags, text, and metadata. The following Figure 1 depicts the varied kinds of similarities examined in this paper.

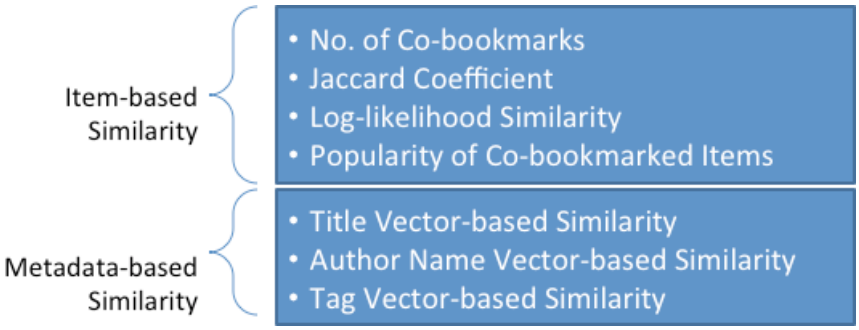


Figure 1. Summary of Similarity Measures

For the rest of this paper, we use the following notation. B is the user-item bookmark matrix, $B = \{B_{ui}\}_{L \times N}$ where L and N denote the number of users and items, respectively. $u \in \{u_1, \ldots, u_l\}$ represents users, and $i \in \{i_1, \ldots, i_n\}$ denotes items. B_i is the bookmarks of item i and B_u is the bookmarks of user u.

4.1 Item-based Similarity

The simplest way to measure information similarity between two users using their bookmark collections is to count the number of items co-bookmarked by both users. This first similarity measure is referred to as the number of co-bookmarks [27]. This is absolute number of common items because we did not consider the proportion of the common items to users' bookmark collections. However, sizes of bookmark collections varied dramatically from user to user. Let's assume that two pairs of users – user a and b; user c and d – have the same number of co-bookmarks, 10. The former pair is richer users, and the union collections of their bookmarks ($B_a \cup B_b$) consist of 250 items. The union collections of the latter pair ($B_c \cup B_d$) have 40 items. For the richer users (a and b), sharing 10 common items is rather trivial, but for the poorer users (user c and d), sharing 10 items means that most of their favorite items are common. We need to measure the number of co-bookmarks using *relative* (normalized) ratios, thereof. As the first relative measure of bookmark-based similarity, we chose the *Jaccard coefficient* [11]. When user a and user b are in consideration, we can compute the coefficient as the following.

$$\text{Jaccard Coefficient } (u_a, u_b) = (B_a \cap B_b)/(B_a \cup B_b) \qquad \text{Eq. 1}$$

The next relative similarity measure is the *log-likelihood similarity*. Log-likelihood similarity was initially proposed as a statistical analysis method for text processing [17]. This method measures how unlikely it is that the overlap of interests occurred simply by chance. The strength of this similarity measure is that it takes into account not only the cases when two people co-bookmarked the same items but also the cases where an item was bookmarked by only one person and even the cases where neither user bookmarked an item [44].

$$\Lambda(u_a, u_b) = \frac{p_{a,b}}{p_b} \times \frac{p_{\sim a,\sim b}}{p_b} = \left(\frac{n_{a,b}}{N_a} \times \frac{N_b}{n_a}\right) \times \left(\frac{n_{\sim a,\sim b}}{N_b} \times \frac{N_a}{n_b}\right) = \frac{n_{a,b} n_{\sim a,\sim b}}{n_a n_b} \qquad \text{Eq.2}$$

where, $\ln(\Lambda(u_a, u_b))$ denotes the log-likelihood similarity between two user a and user b. $n_{a,b}$ is the number of items bookmarked by both user a and b. n_a and n_b is the number of items bookmarked by only user a or user b, respectively. $n_{\sim a,\sim b}$ is the number of items bookmarked by neither of them. A large positive similarity means that two users are very similar and a large negative similarity means high dissimilarity. The similarity value of 0 indicates that two users are neither similar or dissimilar [17].

Due to the popularity bias (i.e., users are more likely to provide feedback on popular items than on rare items [18]), co-bookmarking of a popular item (i.e., a very popular movie like 'Avartar') does not appear to matter as evidence of shared interests as co-bookmarking of a rare item. To account for this, we incorporated the item popularity into information similarity using the average inverse item popularity as referenced below. We refer to this similarity measure as *popularity weight* of co-bookmarked items [40].

$$IP_{ab} = \left(\log \sum_{i \in B_{ab}} \frac{N}{n_i} \right) \Big/ n_{ab} \qquad \text{Eq.3}$$

where IP_{ab} is the popularity weight derived from the average inverse popularity of co-bookmarks between user *a* and *b*. Item *i* is one of the items that the user *a* and *b* commonly bookmarked, and *N* is the total number of items in the dataset. n_i is the number of bookmarks of the item *i* and n_{ab} is the number of co-bookmarks of both user *a* and *b*. A pair of users who shares less popular items thus yields a higher popularity weight value [40] .

4.2 Metadata-Based Similarity

Due to the irregular opportunistic nature of the bookmarking process, users with similar interests may not necessarily end up with similar bookmark collections. In this case, similarity of user interests might be more reliably measured on the level of the item's *metadata.* For example, in a reference management domain such as Citeulike, bookmarking papers from the same authors can be considered as an indicator of similar interests. When titles of two papers in comparison have sufficiently similar keywords, it would be likely that their contents are also similar. In this study, we consider three kinds of textual metadata (paper titles, author names, and social tags) using the vector space model [37]. Even though there are other kinds of metadata associated with bibliographic references such as abstracts and journal/conference names, this data were not available. In particular, we failed to find abstracts and journal/conferences names for more than half the papers on Citeulike. The lack of this data made it impossible to use source-based metadata similarity or topic text similarity measures that require considerable volumes of text.

The first metadata-based similarity is *title vector-based similarity*. All terms in titles of a user's bookmarked papers were aggregated into one bag of keywords. Then, for effective comparison, we applied text-processing techniques to the bag of title keywords. The terms were case-normalized to lower-case letters, and all stop words were removed. We also applied a Porter stemmer to reduce word variation to stems or roots [19]. The processed bag of title keywords was transformed into a title keyword vector consisting of keywords and the TF/IUF (Term Frequency/Inverse User Frequency) values. In this paper, one bag of keywords represents one user's preference model in the aspect of all title keywords in his favorite items. Hence, in our computation, each user is analogous to a document in traditional IDF (Inverse Document Frequency). Therefore, the

inverse user frequency indicates how many users bookmarked papers in which the titles contain a corresponding keyword [15]. Mori, et al. (2006) suggested that, when words in two articles are similar, the contextual representations of the article pair are also similar. In particular, titles of papers succinctly represent the topic of the paper [30]. Therefore, title vector-based similarity tries to compute how much two given users are interested in similar contents.

The second metadata-based similarity uses authors' names, as *author name vector-based similarity*. Authors of academic papers are recognized as experts in specific topics [49]. When a user bookmarks several papers written by an author, it is likely that his interests are related to the author's research topics. Hence, author names are one of the important anchors to determine user interests [49]. Since author names are proper nouns, no text-processing techniques were applied. All author names of one user's bookmarked papers were aggregated as one vector, and the TF/IUF values were computed. Thus, in the same way of building a title keyword vector, the inverse user frequency value of one author name is about how many users bookmarked papers written by the same author.

The last kind of metadata we considered for measuring similarity is social tags. Unlike bookmarks that simply show users' interests in bookmarked items, free-text social tags demonstrate *which aspects* of the tagged items are of interest to the user [15]. Since social tags reflect how each user approaches tagged articles from a personal prospect, it is known that a user's tags represent his conceptual understanding and personal categorization of resources [45]. Thus, using similar tags might be considered as a very strong induction of similar interests.

Before we build a social tag vector, the idiosyncratic nature of social tags requires cleaning up the metadata. In Citeulike, users are able to add free-text tags of their choice on bookmarking items. However, when users do not assign any tags, Citeulike automatically assigns a pseudo-tag, 'no-tag'. When users import bibliographic information from a citation file such RIS, BibIX, ENL files, the system also automatically adds another pseudo-tag such as '*file-import-xx-xx-xx'. These pseudo-tags were added for the system's sake and do not carry any meaning. Therefore, we eliminated the pseudo-tags for the comparison. Then, we built a tag vector using the same text-processing techniques and vector space model used for paper titles (i.e. case normalization, stop-word removal, and stemming)[2]. This is *tag vector-based similarity* [23]. For this tag vector, we computed the TF/IUF values for each tag. The inverse user frequency of this similarity is about how many users annotated their bookmarks with a corresponding tag.

Once all metadata vectors – title vector, author name vector, and processed tag vector – were built, the similarity between a pair of users was computed using the cosine similarity for every metadata vector.

[2] We compared the similarity of original tags without applying any text-processing techniques, and the result showed that the similarity is significantly lower than our proposed similarity based on processed tags. Hence, we omit the similarity measure of original tags.

5. SHARED INTERESTS OF ONLINE SOCIAL RELATOINSHIPS

As explored in the summary Table 1, previous works exploring online homophily have mostly focused on online friendship connections. On the other hand, the homophily of object-centered sociality, such as following and group co-membership connections in social bookmarking systems, has been rarely investigated. The lack of *a priori* grounds of the online object-centered sociality required us to examine whether the two types of the object-centered social connections that we intended to use as gold standard in our analysis have the key property that we expect from a gold standard. In other words, we need to examine whether users connected by watching or co-membership relationship in a social bookmarking system *do* share similar interests. In this section, we perform this analysis. In particular, we test whether users connected by direct relationship (in both watching and group co-membership networks) have a significantly higher interest similarity than those who are indirectly connected or not connected at all.

5.1 Shared Interests of Watching Relations

First, we computed the item-based information similarity according to the users' social distances – direct, 1hop and 2hop – as depicted in Figure 2. To form a baseline, we also computed the similarities of user pairs who are *not connected* with each other by either watching relationship or group co-membership relationship. Because they are not socially associated, and the social reachability is uncountable, we define that the distance between these not connected users is infinite.

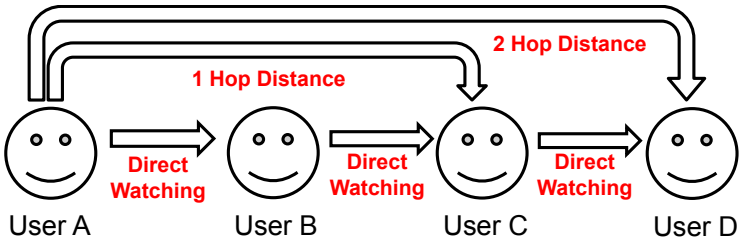


Figure 2. Social Distance (When user A is in direct watching relation with user B, he is in 1hop distance with user C and 2 hop distance with user D.)

As explained in Section 2, the homophily explains that, in the choice of social partners, users make choices through the prism of their information needs and interests, as a process of social comparison or social referrals. Therefore, the information similarity of user pairs in direct watching relationships is expected to be significantly larger than the other pairs in distant relationships, and the similarity values have to be decreased along with the increase of the social distance [46]. Table 3 shows that our presumption is true for the item-based similarity.

For all kinds of item-based measures, user pairs in a direct watching relationship produced the largest degree of similarity, according to the one-way ANOVA test (for number of co-bookmarks, $F = 5192.0$, $p < .001$; for the Jaccard coefficient, $F = 10409.0$, $p < .001$; for popularity, $F = 2396.4$, $p < .001$; and for the log-likelihood similarity, $F = 7091.0$, $p < .001$). In

addition, the similarities were decreased along with the increase of the social distance down to user pairs without any social connection, which have the lowest similarities for all measures. Therefore, we conclude that users connected by a direct watching relationship have the strongest level of homophily. The result of log-likelihood similarity demonstrates that the co-shared items are based on inherently co-shared interests between two users, not by chance [23]. Last, the result about the popularity of co-bookmarked items showed that users connected by direct relationship also tend to share rare items, compared with other users in distant relations or no relation who co-bookmarked rather popular items.

We also compared the metadata-based similarity by the social distances (refer to Table 3). For all kinds of metadata-based measures, the similarities of direct watching relations were the highest and were decreased along with the increase of the social distance, as well (for title vector-based similarity, $F = 21296.7$, $p < .001$; for the author name-based similarity, $F = 8623.1$, $p < .001$; and for tag-based similarity, $F = 19152.0$, $p < .001$). Users in direct relations shared items having the most similar topics and the most similar tag sets than user pairs in distant or infinite social distances. In conclusion, we can identify this result as clear evidence regarding the homophily existing on these watching relations and that the watching relationships are rooted on genuinely common interests. For more detailed results of information similarity comparison in this watching relationship, refer to [21, pp. 107 ~ 132].

Table 3. Comparison of Various Similarity Measures Depending on the Distances of Watching Relations (the values in the last column represent the statistical significance of the comparison)

		Direct	1Hop	2Hop	No Relation	
Item-based Similarities	**No. of Co-bookmarks**	1.80	0.39	0.16	0.04	$F = 5192.0$, $p < .001$
	Jaccard	0.21%	0.04%	0.02%	0.02%	$F = 10408.9$, $p < .001$
	Popularity	8.69	7.75	7.38	6.92	$F = 2396.4$, $p < .001$
	Log_Likelihood	20.4%	9.7%	6.1%	2.3%	$F = 7091.0$, $p < .001$
Metadata-based Similarities	**Title Vector**	14.4%	8.1%	6.3%	1.5%	$F = 21296.7$, $p < .001$
	Author Name Vector	1.5%	0.3%	0.2%	0.07%	$F = 8623.1$, $p < .001$
	Tag Vector	5.1%	1.7%	1.1%	0.2%	$F = 19152.0$, $p < .001$

5.2 Shared Interests of Group Co-Members

Next, we test whether online group co-membership is a sufficiently reliable proxy to represent shared interests among users. Unlike measurable social distance in watching relations, group-based sociability does not have any distance among the members. Therefore, the matter of importance is as to whether co-member pairs in the same group share more similar interests than the other pairs in no relations (neither in direct or indirect watching relations). Table 4 shows the results of the similarity comparison.

Table 4. Various Similarities of Group Co-members in Comparison with the Other User Pairs in No Social Relation

	No. of Co-bookmarks	Jaccard	Popularity	Log Likelihood	Title	Author Name	Tag
Group Co-Members	.26	1.01%	8.00	5.0%	11.2%	2.2%	6.0%
No Relation	.04	0.02%	6.92	2.3%	1.5%	0.07%	0.2%

The item-based similarities of group co-member pairs were compared with the similarity of the other pairs in no social relations. For all kinds of item-based measures, the similarities of co-member pairs were significantly higher than the similarity of the pairs in no social relations ($t = 32.22$, $p < .001$ for the number of co-bookmarks; $t = 58.62$, $p < .001$ for the Jaccard coefficient; $t = 32.24$, $p < .001$ for the log-likelihood similarity; $t = 54.54$, $p < .001$ for the popularity). We also compared metadata-based similarities between two kinds of user pairs. The results of the metadata-based similarities showed the same pattern with the item-based similarities. For all kinds of metadata-based measures, the similarities of group co-members were significantly higher than the similarities of pairs that had no social association ($t = 53.69$, $p < .001$ for the title vector-based similarity; $t = 322.47$, $p < .001$ for author name vector-based similarity; $t = 165.71$, $p < .001$ for tag vector-based similarity). To conclude, apparently, because group co-members shared more similar items, and because their topics of interests are also more similar than the user pairs without any social connection, group co-members have the nature of information-oriented homophily.

6. EFFECTIVENESS OF SIMILARITY MEASURES

In the previous section, we verified that user self-established watching relations and group co-memberships follow the expected patterns of homophily. It provides sufficient evidence to use watching and co-membership relationships as gold standards in this study to assess the quality of various user similarity metrics. This section focuses on the main purpose of this study: comparing the effectiveness of various similarity measures to determine the best performing measure by its ability to predict the existing relationships. The comparative analysis of similarity measures will be examined for each kind of social network in Sections 6.2 and 6.3, respectively.

6.1 Comparison of Similarity Measures

For the comparative analysis of various similarity measures, we borrowed evaluation methods used for personalized recommendation systems. A common way to assess the quality of a specific recommendation approach is to use a subset of user items as a *ground truth* and calculate how well each recommendation approach predicts the ground truth items [8] In our case, different similarity measures can be considered as recommendation approaches that suggest most relevant connections to target users, and the existing Citeulike watching relations and group co-memberships play the role of the ground truth items. Users who participated in either watching network or group memberships are our targets, and we computed the quality

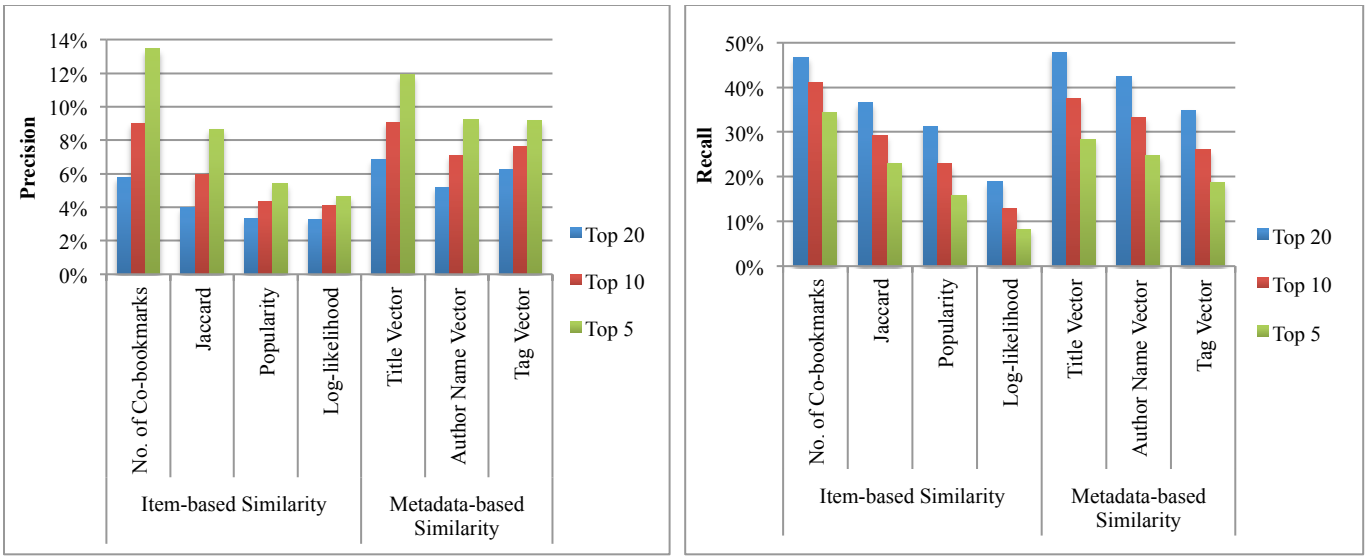
of a specific similarity approach according to its ability to predict "true" social connections of the target users. In our data set, there were 10,114 target users, including 3,223 watching users and 8,009 group members. We computed the similarity of the target users with every other user using all seven kinds of similarity measures (Figure 1), regardless of whether the users are indeed associated. Then, for each measure, we pick the top N similar peers of every target user as recommendations and, among the top N peers, counted how many users were actually socially associated with him or her [17]. We considered three values of N for top-N evaluation – top 20, top 10 and top 5. Under the precondition of homophily, as evidenced by the section 5, we reckon that the more similar interests two users commonly share, the more likely they are socially associated with each other. Using the incremental rank orders, we seek to determine the similarity measure to meet this condition. Precision and recall are used to evaluate the quality of recommendations. Precision assesses the accuracy of the predictions, and recall measures the completeness of the predictions [38]. Specifically, precision at point N (precision@N) is the fraction of correctly predicted ground truth items in the top-N list (Eq. 4). Recall at point N (recall@N) is the fraction of the correctly predicted items in the top-N list to the total number of ground truth items, (i.e., the target users' existing social connections (Eq. 5)).

$$\text{precision@N} = \frac{\text{No. of correct prediction}}{\text{N of top N set}} = \frac{\text{ground truth } \cap \text{top N}}{\text{N}} \quad \text{Eq. 4}$$

$$\text{recall@N} = \frac{\text{No. of correct prediction}}{\text{size of ground truth}} = \frac{\text{ground truth } \cap \text{top N}}{\text{ground truth}} \quad \text{Eq. 5}$$

6.2 Comparing Similarity Measures Using Watching Relations as Ground Truth

First, for 3,223 target users who participated in the watching network, we tested the effectiveness of individual measures by predicting each target user's watched partners. The Figure 3 shows the precision and recall of the predictions made by each similarity measure, respectively. According to the one-way ANOVA test, in the results of both precision and recall, the number of co-bookmarks and title vector-based similarity produced significantly more accurate and complete predictions than the others ($F = 92.0$, $p < .001$ for Top 20 precision; $F = 100.2$, $p < .001$ for Top 10 precision; and $F = 127.9$, $p < .001$ for Top 5 precision; $F = 28.0$, $p < .001$ for Top 20 recall; $F = 20.4$, $p < .001$ for Top 10 recall; and $F = 17.3$, $p < .001$ for Top 5 recall). Between these two measures, there was no significant difference. On the other side, the Jaccard coefficient and the popularity weight were significantly worse than the remaining metadata-based similarities (tag vector-based and author name-based). The poor performance of these item-based measures – Jaccard coefficient, log-likelihood and popularity weight – was surprising, especially in contrast with the excellent performance of a simpler absolute measure. These similarity measures have been widely used in personalized recommendation studies and information retrieval [36, 45]. Note that, in a social bookmarking context, these measures did not perform as we expected.



(**a**) Precisions of Individual Similarity Measures **(b)** Recalls of Individual Similarity Measures

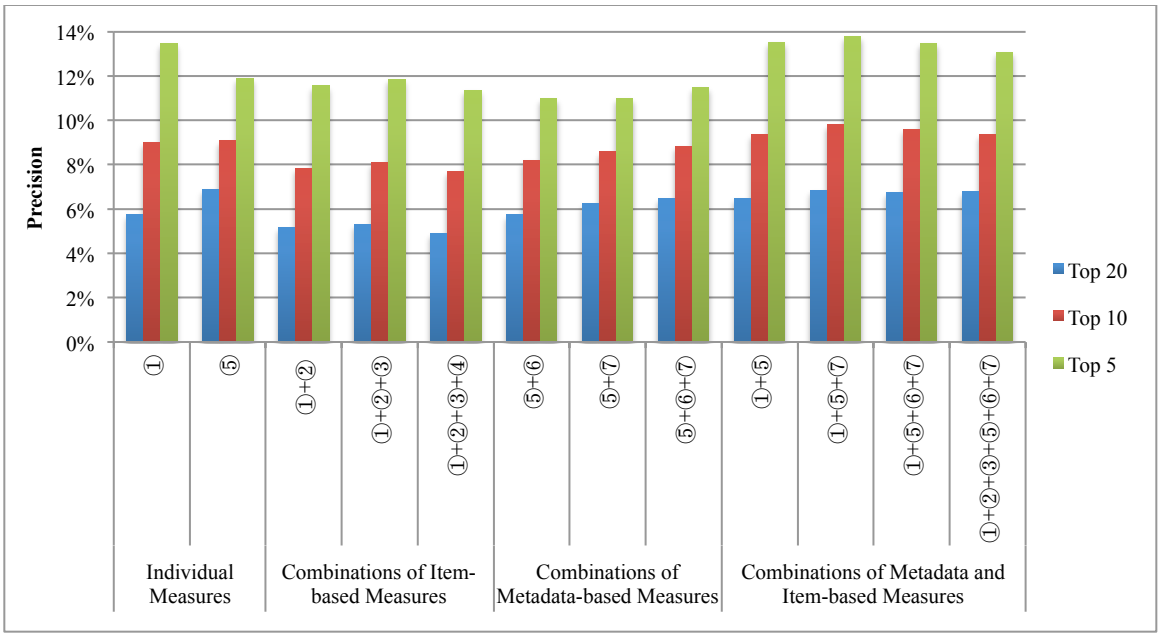
Figure 3. Comparison of Individual Similarity Measures for Predicting Watching Relations

We also examined whether combining multiple measures can increase the capability to represent shared interests among users. In order to combine multiple similarity measures, we need to make them comparable though normalization. Each similarity measure has a heterogeneous range of values. For instance, the number of co-bookmarks could range from 0 to 3,210,960 (i.e. total number of items in our dataset), and Jaccard coefficients and all metadata-based similarities range from 0 to 1. Hence, before fusing similarity measures together, we normalized each similarity measure using the Standard Score (SS) like the following.

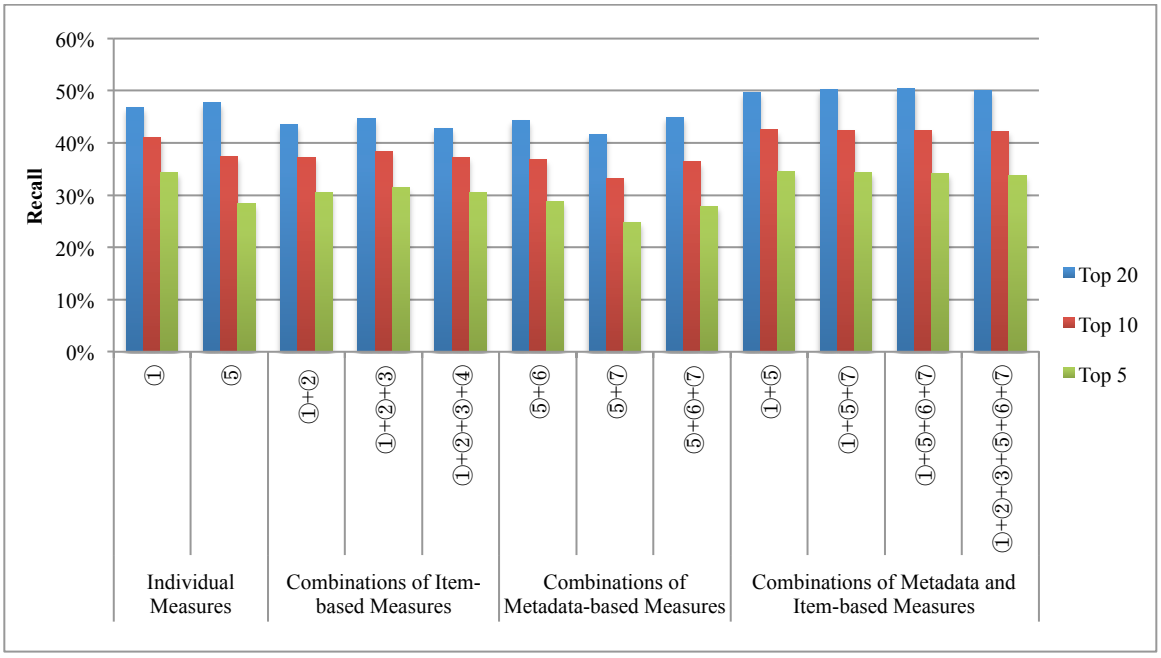
$$SS = \frac{x - \mu}{\sigma}$$

where x is one similarity value, and μ and σ is the mean and standard deviation of the corresponding similarity measure, respectively. Then, we combined the normalized similarity values together by summing up the normalized similarity values with equal weights and predicted existing watching relations using the combined similarities. Figure 4 shows the results of precisions and recalls for the combined measures along with the best two individual similarities (i.e. the number of co-bookmarks and title vector-based similarity).

In the precision results, we found that, at all three top N ranks, the combination of the number of co-bookmarks, title vector-based similarity, and tag vector-based similarity (i.e. ①+⑤+⑥ on the Figure 4(a)) significantly outperformed all other similarities ($F = 57.54$, $p < .001$ for top 20, $F = 65.49$, $p < .001$ for top 10, $F = 83.84$, $p < .001$ for top 5). Moreover, the post-hoc test concluded that the precision of the best-performed combination is significantly better than the precisions of the best individual similarity measures (i.e. the number of co-bookmarks).



(a) Precisions of Combined Measures



(b) Recalls of Combined Measures

Figure 4. Comparison of Combined Measures for Predicting Watching Relations (① = No. of Co-bookmarks; ② = Jaccard; ③ = Popularity; ④ = Log-likelihood; ⑤ = Title Vector; ⑥ = Author Name Vector; ⑦ = Tag Vector; The data series is displayed in the order of the legend)

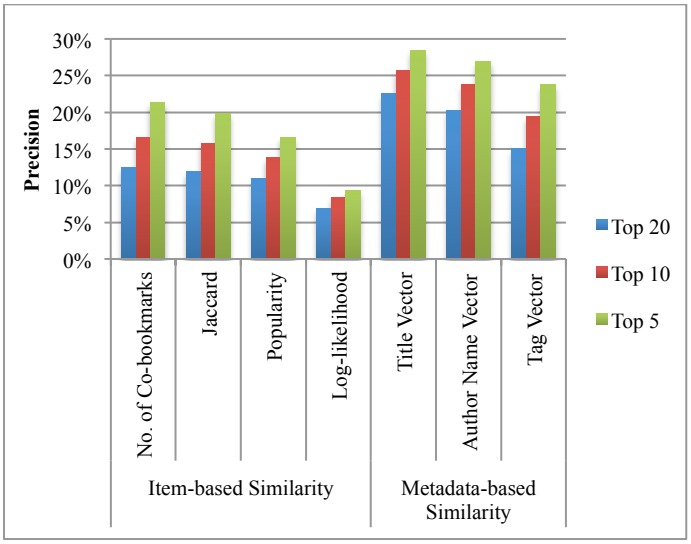
The one-way ANOVA test on the recall results showed that, at higher rank (i.e. top 5 and 10), two kinds of combinations – the combinations of the number of co-bookmarks and title vector-based similarity (i.e. ①+⑤ on the Figure 4(b)) and the combinations of the; number of co-bookmarks, title vector-based similarity and tag vector-based similarity (i.e. ①+⑤+⑦ on the Figure 4(b)) – significantly outperformed other combined similarity measures ($F = 94.20$, $p < .001$ for top 20, $F = 101.04$, $p < .001$ for top 10, $F = 96.98$, $p < .001$ for top 5). These two kinds of combinations were not significantly different.

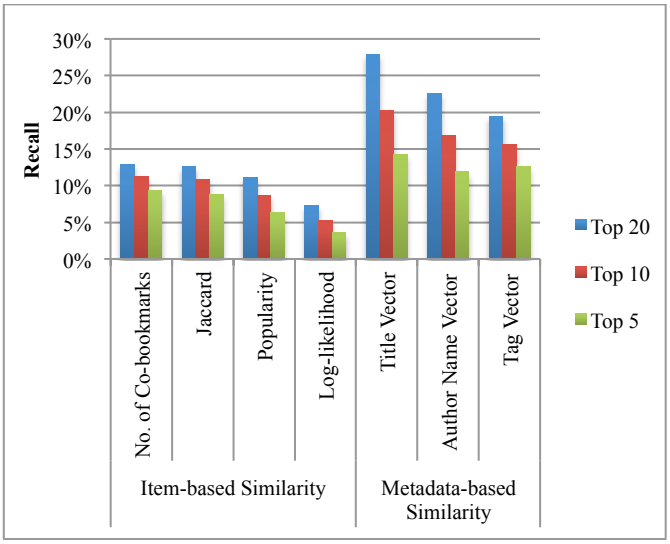
Equal to the result of precision, the recalls of the best-performed combinations were significantly higher than two best individual measures (i.e. the number of co-bookmarks and title vector-based similarity). In Figure 4, we omitted the results of the other combinations, when they yielded significantly lower precisions and recalls.

To conclude, the results of both evaluation criteria demonstrated that the most critical information to represent the common interests of watching relations is how many bookmarks they share, what the bookmarked items are about and which tags they use to annotate their bookmarked items.

6.3 Comparing Similarity Measures Using Group Co-Membership as Ground Truth

To compare similarity measures using group co-membership data, we chose 8,009 target users who were members of group(s) and attempted to predict their co-memberships of the same groups using each of the explored similarity measures. Figure 5 shows the precision and recall of the predictions made by individual similarity measures. In all top N lists and for both evaluation criteria, all kinds of metadata-based measures significantly outperformed the item-based measures ($F = 329.9$, $p < .001$ for top 20 precision; $F = 305.9$, $p < .001$ for top 10 precision; $F = 289.5$, $p < .001$ for top 5 precision; $F = 472.2$, $p < .001$ for top 20 recall; $F = 279.4$, $p < .001$ for top 10 recall; $F = 174.5$, $p < .001$ for top 5 recall). Even the number of co-bookmarks, which was the best individual similarity measure for watching connections, performed worse than any of the metadata-based similarities for the group co-member relationships. Among metadata-based similarities, title vector-based similarity provided the best prediction of co-membership in terms of both precision and recall with the statistical significance at all three ranks. This result implies, as discussed in Section 5, the group co-members frequently bookmark items that are conceptually and semantically similar, yet less-frequently bookmark exactly the same items. Hence, group co-members conclusively have common interests and similar conceptualization of information items. This is why the measures based on co-bookmarking identical items performed much worse than metadata-based measures.

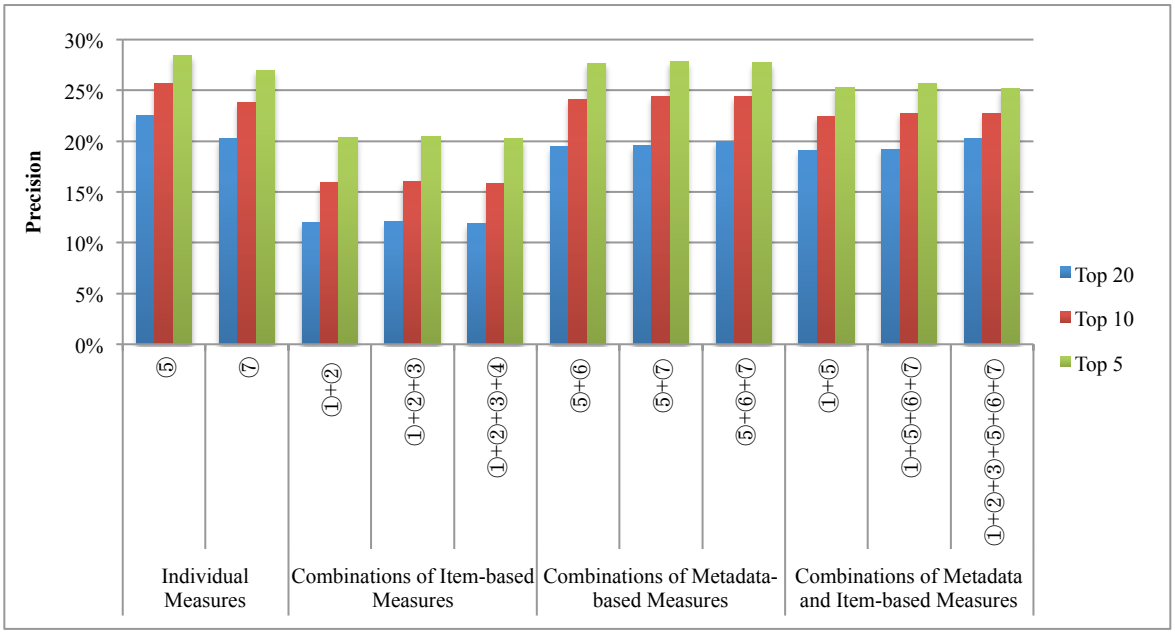




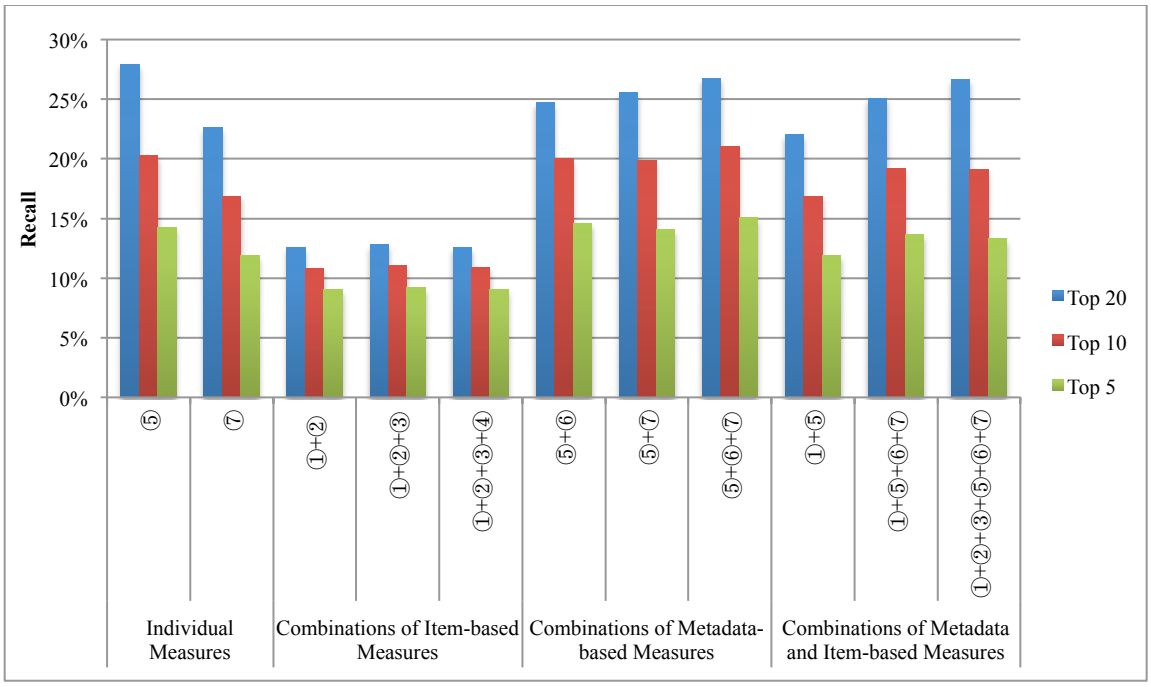
(**a**) Precisions of Individual Similarity Measures (**b**) Recalls of Individual Similarity Measures

Figure 5. Comparison of Individual Similarity Measures for Predicting Group Co-members

We also evaluated the performance of various combinations of similarities by predicting ground truth group co-members. Figure 6 shows the results of precision and recall along with the two best individual measures. In the precision analysis, naturally, the combinations of metadata-based similarities performed significantly better than the combinations consisting solely of item-based similarities ($F = 136.9$, $p < .001$ for the top 20; $F = 116.1$, $p < .001$ for the top 10; $F = 70.3$, $p < .001$ for the top 5). There were no significant differences among the varied combinations of metadata-based similarities (⑤+⑥, ⑤+⑦, and ⑤+⑥+⑦ in the Figure 6(a)). The interesting pattern that we found here is that the best-performing individual similarity – title vector-based similarity – produced as accurate predictions as did these combinations of metadata-based similarities. For example, between the title vector-based similarity and one of the best-performing combinations of metadata (i.e. ⑤+⑥+⑦ in Figure 6(a)), the precision values were not significantly different ($t = 2.4$, $p = 0.021$ for the top 20; $t = 1.9$, $p = 0.10$ for the top 10; $t = 1.4$, $p = 0.13$ for the top 5). In the recall analysis, the combinations of all metadata-based similarities (i.e. ⑤+⑥+⑦ in Figure 6(b)) also performed significantly better than the other combinations of item-based similarities ($F = 283.0$, $p < .001$ for the top 20; $F = 156.9$, $p < .001$ for the top 10; $F = 67.2$, $p < .001$ for the top 5). In here, like. Here, as with the preceding result of precision, the most notable result is that the best individual measure (title vector-based similarity) predicted as many complete predictions as did the best combinations at all ranks without any significant statistical difference. Consequently, various combinations of similarities did not add significant improvement to this single best measure, which is based on titles.



(a) Precisions of Combined Measures



(b) Recall of Combined Measures

Figure 6. Comparison of Combined Measures for Predicting Group Co-members (①= No. of Co-bookmarks; ② = Jaccard; ③= Popularity; ④ = Log-likelihood; ⑤ = Title Vector; ⑥ = Author Name Vector; ⑦ = Tag Vector; The data series is displayed in the order of the legend)

In conclusion, in order to represent the shared interests among group co-members, the metadata-based similarity measures are significantly more effective than item-based similarity measures. In addition, for group co-membership, combining multiple similarity measures could not intensify the effectiveness of the best individual measure: title keywords of bookmarked items appeared to be the most critical information in which to express shared interests of group members.

7. CONCLUSION AND DISCUSSION

The main purpose of this study was to determine the most effective similarity measure for estimating similarity of user interests in object-centered online social networks. Finding the best similarity measure is important not only academically but also practically because estimating shared user interests is the key component for two popular recommendation targets: recommending like-minded peers and recommending relevant items. As the context for our study, we used Citeulike, one of the most popular social bookmark systems. A critical value of Citeulike in the context of our study is the presence of a considerable volume of not just one, but two types of self-defined object-centered connections between users: 1) *watching* relationship and 2) co-membership relationship. These self-defined connections are highly valuable as a golden standard to compare various similarity measures. We started our analysis with asserting the gold-standard status of these social connections by confirming the *homophily hypothesis* for both types of Citeulike social connections. After that, we compared

several item-based and metadata-based similarity measures for the two types of self-defined connections using the evaluation concepts and methodologies from personalized recommendation research. We assessed the performance of seven similarity measures according to the ability to predict the existence of self-defined connections.

Our analysis brought interesting results. First, metadata-based similarity measures, specifically title-based similarity in our context, perform generally better than item-based similarity measures. For watching relationships, the absolute number of common bookmarks – one of the popular item-based similarity measures – competed with metadata-based measures. However, when we combined the titles with other two types of metadata (i.e. authors and tags), the combined metadata similarity measure performed better than the number of common bookmarks. Moreover, for group co-memberships, the title-based similarity measures consistently outperformed all other similarity measures. Hence, the title vector-based similarity emerged as the most effective solo measure that worked well for both watching relationships and group co-memberships.

This result is important from both practical and methodological prospects. In traditional recommendation research, item-based similarity approaches based on ratings are considered reliable and dominant (refer to the section 2.1). Our work, however, suggests that, instead of blindly applying a similarity measure that performed well in one type of system to another type of system, it is wiser to compare the performance of a range of potential measures in the new context. In particular, in our context, a combination of metadata-based similarity approaches emerged as a better candidate to serve as a core of a user recommender system. While this result might look surprising, we believe that it might be explained by the nature of social bookmarking. In a social bookmarking system such as Citeulike, the ratio of the number of items to the number of users is typically much larger than in product recommender systems. In this context, users have a much lower chance to bookmark a sufficient number of the *same items*, which makes all item-based similarity measures less reliable. As a result, rather than expecting users to bookmark a sufficient number of the exactly same papers to be considered similar, it is wiser to look at the metadata of bookmarked papers (especially types of metadata to represent contents succinctly like titles) for a more reliable similarity estimate.

We also discovered that the absolute number of shared bookmarks was able to successfully compete with otherwise best performing metadata-based measures when watching relationship was used as the gold standard, while it was left significantly behind in the race with metadata measures for group co-membership relationship. We believe this is the result of the specific nature of the watching relationship in Citeulike: once established, the social dynamic encourages copying items from the watched party. It increases the chances of sharing bookmarks and makes the shared bookmarks a natural “reverse

predictor" of watching. The much better prediction performance of the simpler approach to count shared bookmarks over more elaborated relative measures (i.e. Jaccard coefficient, log-likelihood similarity and popularity weights) provides us with supporting evidence about the nature of watching relationships centered on objects of interests.

The above observations point to two directions of future work. One is to use other types of gold standards to compare similarity measures. In particular, we are interested in comparing the performance of other self-defined social relationships from other social bookmarking systems or social networking sites. Another direction of future work is to explore other types of similarity measures. This paper focused on most popular similarity measures since they are already recognized and trusted and thus are most likely candidates for adoption. Our next target is a set of more advanced approaches that are based on graph structures of social networks and use more sophisticated machine learning techniques. Throughout this study, we have emphasized that the current social structures are strongly tied to knowledge-driven online culture. The online social relations in our consideration met the similarity attraction hypothesis [16] and held the transitivity power of information [3]; however, we did not consider any social graph structure-based similarity measures. In [26], Ma demonstrated that information similarity varies according to the social properties such as the number of shared friends, subgraph topology and connected components. In the future, we plan to include social structure-based similarity measures with the given assumption that the social closeness could represent the resemblance of information interests. In addition, we plan to engage richer datasets to explore more computationally sophisticated similarity measures based on the dimensionality reduction and rooted in semantic and lexical structures of textual metadata. Lastly, instead of more traditional similarity fusion with equal weights, we plan to explore statistics-based fusion approaches with fusion weight inferred from data.

REFERENCES

[1] Akcora, C.G., B. Carminati, and E. Ferrari. *Network and profile based measures for user similarities on social networks*. in *Proceedings of IEEE International Conference on Information Reuse and Integration (IRI)*. 2011.

[2] Anderson, A., et al., *Effects of user similarity in social media*, in *Proceedings of the fifth ACM international conference on Web search and data mining*. 2012, ACM: Seattle, Washington, USA. p. 703-712.

[3] Antoniades, D. and C. Dovrolis, *Co-evolutionary dynamics in social networks: A case study of Twitter*. Computational Social Networks, 2015. **2**(1): p. 1.

[4] Baatarjav, E.-A., S. Phithakkitnukoon, and R. Dantu, *Group Recommendation System for Facebook*, in *Proceedings of the OTM Confederated International Workshops and Posters on On the Move to Meaningful Internet Systems: 2008 Workshops: ADI, AWeSoMe, COMBEK, EI2N, IWSSA, MONET, OnToContent + QSI, ORM, PerSys, RDDS, SEMELS, and SWWS*. 2008, Springer-Verlag: Monterrey, Mexico. p. 211-219.

[5] Bakshy, E., et al., *The role of social networks in information diffusion*, in *Proceedings of the 21st international conference on World Wide Web*. 2012, ACM: Lyon, France. p. 519-528.

[6] Bhattacharyya, P., A. Garg, and S. Wu, *Analysis of user keyword similarity in online social networks*. Social Network Analysis and Mining, 2011. **1**(3): p. 143-158.

[7] Bischoff, K., *We love rock 'n' roll: analyzing and predicting friendship links in Last.fm*, in *Proceedings of the 3rd Annual ACM Web Science Conference*. 2012, ACM: Evanston, Illinois. p. 47-56.

[8] Bobadilla, J., et al., *Recommender systems survey*. Knowledge-Based Systems, 2013. **46**(0): p. 109-132.

[9] Boj, U., et al., *Interlinking the Social Web with Semantics*. IEEE Intelligent Systems, 2008. **23**(3): p. 29-40.

[10] Brzozowski, M.J., T. Hogg, and G. Szabo, *Friends and foes: ideological social networking*, in *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*. 2008, ACM: Florence, Italy. p. 817-820.
[11] Buccafurri, F., et al., *Discovering missing me edges across social networks.* Information Sciences, 2015. **319**: p. 18-37.
[12] Chen, D.-B., et al., *Identifying influential nodes in large-scale directed networks: the role of clustering.* PloS one, 2013. **8**(10): p. e77455.
[13] Colucci, L., et al. *Evaluating Item-Item Similarity Algorithms for Movies*. in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 2016. ACM.
[14] Hajian, B. and T. White. *Modelling Influence in a Social Network: Metrics and Evaluation*. in *Proceedings of IEEE third international conference on Privacy, security, risk and trust (passat) and 2011 ieee third international conference on social computing (socialcom)*. 2011.
[15] Huang, C.-L., et al., *Utilizing user tag-based interests in recommender systems for social resource sharing websites.* Knowledge-Based Systems, 2014. **56**: p. 86-96.
[16] Ilmarinen, V.-J., J.-E. Lönnqvist, and S. Paunonen, *Similarity-attraction effects in friendship formation: Honest platoon-mates prefer each other but dishonest do not.* Personality and Individual Differences, 2016. **92**: p. 153-158.
[17] James, K., et al., *Multiple Gold Standards Address Bias in Functional Network Integration*, in *Newcastle University, Computing Science Technical Reports Series*. 2011, Newcastle University.
[18] Jonnalagedda, N., et al., *Incorporating popularity in a personalized news recommender system.* PeerJ Computer Science, 2016. **2**: p. e63.
[19] Karaa, W.B.A. and N. Gribâa, *Information retrieval with porter stemmer: a new version for English*, in *Advances in Computational Science, Engineering and Information Technology*. 2013, Springer. p. 243-254.
[20] Knoll, L.J., et al., *Social Influence on Risk Perception During Adolescence.* Psychological Science, 2015. **26**(5): p. 583-592.
[21] Lee, D.H., *Personalized Recommendations Based on Users' Information Centered Social Networks*, in *Information Science*. 2013, University of Pittsburgh: Pittsburgh, PA, USA.
[22] Lee, D.H. and T. Schleyer, *Social tagging is no substitute for controlled indexing: A comparison of Medical Subject Headings and CiteULike tags assigned to 231,388 papers.* Journal of the American Society for Information Science and Technology, 2012. **63**(9): p. 1747-1757.
[23] Levy, M. and M. Sandler, *Music Information Retrieval Using Social Tags and Audio.* IEEE Transactions on Multimedia, 2009. **11**(3): p. 383-395.
[24] Liu, H., et al., *A new user similarity model to improve the accuracy of collaborative filtering.* Knowledge-Based Systems, 2014. **56**: p. 156-166.
[25] Lönnqvist, J.-E. and J.V.A. Itkonen, *Homogeneity of personal values and personality traits in Facebook social networks.* Journal of Research in Personality, 2016. **60**: p. 24-35.
[26] Ma, H., *On measuring social friend interest similarities in recommender systems*, in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 2014, ACM: Gold Coast, Queensland, Australia. p. 465-474.
[27] Manca, M., L. Boratto, and S. Carta. *Mining User Behavior in a Social Bookmarking System-A Delicious Friend Recommender System*. in *Proceedings of the 3rd International Conference on Data Management Technologies and Applications*. 2014.
[28] McPherson, M., S. Lovin, and J. Cook, *BIRDS OF A FEATHER: Homophily in Social Networks.* Annual Review of Sociology, 2001. **27**: p. 415-445.
[29] Modani, N., et al., *Like-Minded Communities: Bringing the Familiarity and Similarity together*, in *Proceedings of Web Information Systems Engineering*. 2012. p. 382-395.
[30] Mori, J., et al., *Extracting Relations in Social Networks from the Web Using Similarity Between Collective Contexts*, in *Proceedings of 5th International Semantic Web Conference on the Semantic Web - ISWC 2006*. 2006, Springer Berlin Heidelberg. p. 487-500.
[31] Nansen, B., et al., *An internet of social things*, in *Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: the Future of Design*. 2014, ACM: Sydney, New South Wales, Australia. p. 87-96.
[32] Nicolini, D., J. Mengis, and J. Swan, *Understanding the role of objects in cross-disciplinary collaboration.* Organization Science, 2012. **23**(3): p. 612-629.
[33] Oh, S. and S.Y. Syn, *Motivations for sharing information and social support in social media: A comparative analysis of Facebook, Twitter, Delicious, YouTube, and Flickr.* Journal of the Association for Information Science and Technology, 2015: p. n/a-n/a.
[34] Onodera, N. and F. Yoshikane, *Factors affecting citation rates of research articles.* Journal of the Association for Information Science and Technology, 2015. **66**(4): p. 739-764.

[35] Park, J.H., et al., *An investigation of information sharing and seeking behaviors in online investment communities.* Computers in Human Behavior, 2014. **31**: p. 1-12.
[36] Patel, R. and P. Bhatt, *A survey on semantic focused web crawler for information discovery using data mining technique.* International Journal for Innovative Research in Science and Technology, 2015. **1**(7): p. 168-170.
[37] Saari, P. and T. Eerola, *Semantic computing of moods based on tags in social media of music.* IEEE Transactions on Knowledge and Data Engineering, 2014. **26**(10): p. 2548-2560.
[38] Schafer, 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 Heidelberg: Berlin, Heidelberg. p. 291-324.
[39] Singla, P. and M. Richardson, *Yes, there is a correlation: - from social networks to personal behavior on the web*, in *Proceeding of the 17th international conference on World Wide Web*. 2008, ACM: Beijing, China. p. 655-664.
[40] Steck, H., *Item popularity and recommendation accuracy*, in *Proceedings of the fifth ACM conference on Recommender systems*. 2011, ACM: Chicago, Illinois, USA. p. 125-132.
[41] Subramaniyaswamy, V., V. Vijayakumar, and V. Indragandhi, *A Review of Ontology-Based Tag Recommendation Approaches.* International Journal of Intelligent Systems, 2013. **28**(11): p. 1054-1071.
[42] Tapia-Rosero, A., A. Bronselaer, and G. De Tré, *A method based on shape-similarity for detecting similar opinions in group decision-making.* Information Sciences, 2014. **258**: p. 291-311.
[43] Weng, L., et al., *The role of information diffusion in the evolution of social networks*, in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2013, ACM: Chicago, Illinois, USA. p. 356-364.
[44] Wilson, J., S. Chaudhury, and B. Lall. *Improving Collaborative Filtering based Recommenders using Topic Modelling*. in *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 01*. 2014. IEEE Computer Society.
[45] Xie, H., et al., *Incorporating sentiment into tag-based user profiles and resource profiles for personalized search in folksonomy.* Information Processing & Management, 2016. **52**(1): p. 61-72.
[46] Yavaş, M. and G. Yücel, *Impact of Homophily on Diffusion Dynamics Over Social Networks.* Soc. Sci. Comput. Rev., 2014. **32**(3): p. 354-372.
[47] Ylijoki, O. and J. Porras, *Conceptualizing Big Data: Analysis of Case Studies.* Intelligent Systems in Accounting, Finance and Management, 2016. **23**(4): p. 295-310.
[48] Yu, Y., L. Mo, and J. Zhou, *Social Friend Interest Similarity in Microblog and its Implication.* International Journal of Control and Automation, 2015. **8**(11): p. 21-32.
[49] Zhao, D. and A. Strotmann, *The knowledge base and research front of information science 2006–2010: An author cocitation and bibliographic coupling analysis.* Journal of the Association for Information Science and Technology, 2014. **65**(5): p. 995-1006.
[50] Zuo, X. and A. Iamnitchi, *A Survey of Socially Aware Peer-to-Peer Systems.* ACM Comput. Surv., 2016. **49**(1): p. 1-28.