Information Sciences 418–419 (2017) 46–60

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

How to measure information similarity in online social networks: A case study of Citeulike

Danielle H. Lee [a,*], Peter Brusilovsky [b]

[a] Adaptive Interactions, Co., 1460-1 Bangbae-dong, Seocho-gu, Seoul, 06568, South Korea
[b] School of Information Sciences, University of Pittsburgh. 135 N. Bellefield Ave., Pittsburgh PA., 15260, United States

ARTICLE INFO

Article history:
Received 21 July 2015
Revised 11 April 2017
Accepted 29 July 2017
Available online 31 July 2017

Keywords:
Information similarity
Online social networks
Watching relations
Group membership
Social tags
Citeulike

ABSTRACT

In our current knowledge-driven society, many information systems encourage users to utilize their online social connections' information collections actively 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 measure effectively 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 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 shared interests.

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

* Corresponding author at: Computing and Software Systems, 14170 NE 183rd Street, Woodinville, WA, 98072, United States.
E-mail addresses: suleehs@gmail.com (D.H. Lee), peterb@pitt.edu (P. Brusilovsky).

http://dx.doi.org/10.1016/j.ins.2017.07.034
0020-0255/

to 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. This 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 users and compute item similarity among users solely based on item ratings to identify like-minded peers. However, most social media systems provide no rating function. 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 as indicated by ratings. 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 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, as other existing studies have done? 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 [35]. Thus, the presence of the same authors or keywords in the collections of two users could indeed indicate overlapping research interests between the users.

Focusing on the number of common items between users without any consideration of content or metadata is the closest analogy to the current rating-based approach in collaborative filtering, and was the first to be explored [14]. However, it has been also argued that content-oriented similarity approaches are more meaningful for social systems in which 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 [42]. As a result of the large information flow, like-minded users tend to have a much lower chance of sharing or adopting exactly the 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 [37], no systematic evaluation and comparison of these approaches has yet been performed. To bridge this gap, this paper aims to analyze comparatively 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 the social associations are likely predicated on target users' perceived similarity to truly like-minded users and the ensuing information utility. Taken altogether, the contribution of our paper is to execute a systematic and comparative evaluation of several popular similarity measures in circumstances where users' ratings are unavailable and users' preferences are expressed in multifarious formats. Through an examination of candidate gold standards for measuring user similarity in social systems, we will find an optimal similarity measure that could be used to identify genuinely like-minded users in social systems.

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 with examining the properties of these connections to confirm that they could indeed serve as gold standards. Once this has been 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).

2. Related works

2.1. Online social networks and interest similarity

Table 1 summarizes existing studies that have 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