Information Sciences 319 (2015) 18–37

Contents lists available at ScienceDirect

Information Sciences

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

Discovering missing me edges across social networks ☆

Francesco Buccafurri *, Gianluca Lax, Antonino Nocera, Domenico Ursino

DIIES, Università Mediterranea di Reggio Calabria, Via Graziella, Località Feo di Vito, 89122 Reggio Calabria, Italy

ARTICLE INFO

Article history:
Received 26 February 2014
Received in revised form 2 February 2015
Accepted 1 May 2015
Available online 18 May 2015

Keywords:
Social networks
Identity management
Membership overlap

ABSTRACT

Distinct social networks are interconnected via membership overlap, which plays a key role when crossing information is investigated in the context of multiple-social-network analysis. Unfortunately, users do not always make their membership to two distinct social networks explicit, by specifying the so-called me edge (practically, corresponding to a link between the two accounts), thus missing a potentially very useful information. As a consequence, discovering missing me edges is an important problem to address in this context with potential powerful applications. In this paper, we propose a common-neighbor approach to detecting missing me edges, which returns good results in real-life settings. Indeed, an experimental campaign shows both that the state-of-the-art common-neighbor approaches cannot be effectively applied to our problem and, conversely, that our approach returns precise and complete results.

1. Introduction

In recent years, (on-line) social networks have been showing a rapid development growth becoming probably the main actor of the Web 2.0. The rapid and revolutionary diffusion of social networks among all segments of the population has attracted the interest of researchers from several fields of computer science, such as digital forensics [40,24], user behavior [6], trust and reputation [26], steganography [25,28], also for the applications that the analysis of involved data can enable [61,22,13,38,14,12]. In this landscape, Social Network Analysis and Social Network Mining have assumed an important role because both the large volume of data and their graph-based organization have enforced the development of specific models and methods allowing the study of social-network data to discover knowledge from them. Clearly, the graph-based data schema gives a great information power to links among data, because it allows people profiles, resources, activities, and so on, to be directly (and indirectly) related. The crucial role of relationships in the expression of an individual's personality and social identity, traditionally recognized by social sciences, is even strengthened in the field of virtual societies, in which relationship links are the main form of expression of participation of individuals to the community. To make more challenging the analysis of this reality, consider that the reference scenario does not consist of a single, isolated, independent social network, but is a constellation of social networks, each forming a community with specific connotations, but strongly interconnected with each other. It is a matter of fact that, despite the inherent underlying heterogeneity, the interaction among distinct social networks is the basis of a new emergent internetworking scenario enabling a lot of strategic applications,

☆ A shorter abridged version of this paper appears in "Discovering Links among Social Networks", by F. Buccafurri, G. Lax, A. Nocera, and D. Ursino, in the Proc. of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2012), Bristol, United Kingdom, 2012. Springer [17].

* Corresponding author.
E-mail addresses: bucca@unirc.it (F. Buccafurri), lax@unirc.it (G. Lax), a.nocera@unirc.it (A. Nocera), ursino@unirc.it (D. Ursino).

http://dx.doi.org/10.1016/j.ins.2015.05.014
0020-0255/

whose main strength will be just the integration of possibly different communities yet preserving their diversity and autonomy. Clearly, social mining and analysis approaches may strongly rely on this huge multi-network source of information, which reflects multiple aspects of people personal life, thus enabling a lot of powerful discovering activities.

From this perspective, links among different social networks assume a fundamental role. They connect the same user on two different social networks who thus assumes the role of passing point of information from one social network the other. For this reason, we call this user *i-bridge*.[1] The link derives from the explicit user's declaration (sometimes supported and encouraged by specific tools) consisting in the insertion of `me` edges [1]. Unfortunately, for disparate reasons, users do not always make their membership to two distinct social networks explicit, by specifying the so-called `me` edge (practically corresponding to a link between the two accounts), thus missing a potentially very useful information. As a consequence, in the overall underlying (social internetworking) graph a big number of missed `me` edges exists, whose discovery represents a very important issue. In other words, an interesting problem of missing link detection arises, which partially overlaps with a link prediction issue, because we may expect that a portion of missing `me` edges will be inserted in a next stage in the graph.

In this paper, we deal with the above problem by proposing an effective solution experimentally tested in a real-life Social Internetworking Scenario (SIS, for short) [15]. To the best of our knowledge, the problem of detecting `me` edges has not been investigated in the literature, but the approach we adopt in this work, which exploits a recursive notion of common-neighbor similarity, suggested us to prior verify whether common-neighbor approaches for link prediction [47] can be directly applied to our problem. The answer to this question was definitely negative, as intuitively explained in Section 2 and experimentally confirmed in Section 6, thus motivating our work. Our solution is based on a notion of node similarity, whose usage allows us to detect whether a suitable threshold is exceeded and then a missing `me` edge between two nodes is detected. The similarity between two nodes is obtained by combining two contributions: a string similarity between the associated usernames, and a contribution based on a suitable recursive notion of common-neighbor similarity. The neighborhood similarity allows these errors to be detected and avoided. As a consequence, it is important to clarify that the problem we are addressing does not deal with the case in which a user with membership overlap between two social networks chooses the corresponding account names very different from each other. It is worth noting that, under this case, often falls the situation in which a user voluntarily keeps the two accounts separated in their respective social networks and thus avoid also to have common friends. Therefore, also neighborhood similarity fails.

The plan of this paper is as follows: in the next section, we examine related literature. In Section 3, we present our recursive notion of similarity. On the basis of this notion, we design the method we use to detect missing `me` edges. This is described in Section 4. In Section 5, we determine the computational complexity of our approach. In Section 6, we illustrate the experiments we have carried out to verify the performances of our technique. Finally, in Section 7, we draw our conclusions and sketch possible future evolutions of our research.

2. Related work

In this section, we survey the literature related to the topics addressed in this paper. It is discussed subdivided into three categories, one for each subsection.

2.1. Identifying users on the Web

The detection of `me` edges in a SIS can be seen as a special case of the problem of identifying users on the Web. As a matter of fact, it allows the features of i-bridge users to be detected. Identifying users on the Web has received a great attention in several application scenarios, such as personalization. A lot of work is devoted to verify whether user profile information can be sufficient to address this problem. In [23] the authors define and implement a framework that provides a common base for user identification for cross-system personalization among Web-based user-adaptive systems. The corresponding user identification algorithm combines a set of identification properties, such as username, name, location or email address, and classifies a user as identified if such a combination exceeds a suitable threshold. In [42], a technique based on user profiles for identifying users across social systems is proposed. This technique has been successfully validated on three social tagging networks (Flickr, Delicious and StumbleUpon). The limit of this technique is that only few users make their profile available in social tagging platforms. A method to identify users on the basis of profile matching is proposed in [60]. In this paper, data from two popular social networks are used to evaluate the importance of fields in the Web profile and to develop a profile comparison tool. The authors of [64] provide evidence on the existence of mappings among usernames across different communities. Starting from the observation of the data in BlogCatalog, they infer 7 hypotheses on the relationships among the usernames selected by a single person in different communities. On the basis of such hypotheses, they propose an approach that, given a username u in a source community and a target community c, generates a set of candidate usernames in c corresponding to u. The approach first generates a set of usernames from u by adding and removing suitable prefixes and suffixes. Then, it exploits a Web search on Google aimed at checking for the existence of each candidate username in such a way as to reduce the returned set of usernames.

[1] The prefix "i-" stands for "internetworking" and is used to avoid ambiguity with the classic notion of "bridge" [33]. Observe that an i-bridge is a node, whereas a (classic) bridge is an edge.