Information and Software Technology xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

Information and Software Technology

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

Who to follow recommendation in large-scale online development communities

Daniel Schall *

Siemens Corporate Technology, Siemensstrasse 90, 1211 Vienna, Austria

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Online communities
Software development
GitHub
Knowledge management
Who-to-follow recommendation
Social resilience

ABSTRACT

Context: Open source development allows a large number of people to reuse and contribute source code to the community. Social networking features open opportunities for information discovery, social collaborations, and improved recommendations of potential collaborators.
Objective: Online community and development platforms rely on social network features to increase awareness and attention among community members for improved collaborations. The objective of this work is to introduce an approach for recommending relevant users to follow. Follower networks provide means for informal information propagation. The efficiency and effectiveness of such information flows is impacted by the network structure. Here, we aim to understand the resilience of networks against random or strategic node removal.
Method: Social network features of online software development communities present a new opportunity to enhance online collaboration. Our approach is based on the automatic analysis of user behavior and network structure. The proposed 'who to follow' recommendation algorithm can be parametrized for specific contexts. Link-analysis techniques such as PageRank/HITS provide the basis for a novel 'who to follow' recommendation model.
Results: We tested the approach using a GitHub-based dataset. Currently, users follow popular community members to get updates regarding their activities instead of maintaining personal relations. Thus, social network features require further improvements to increase reciprocity. The application of our 'who to follow' recommendation model using the GitHub dataset shows excellent results with respect to context-sensitive following recommendations. The sensitivity of GitHub's follower network to random node removal is comparable with other social networks but more sensitive to follower authority based node removal.
Conclusion: Link-based algorithm can be used for context-sensitive 'who to follow' recommendations. GitHub is highly sensitive to authority based node removal. Information flow established through follower relations will be strongly impacted if many authorities are removed from the network. This underpins the importance of 'central' users and the validity of focusing the 'who to follow' recommendations on those users.

1. Introduction

Social networks have become a central part for many people in their everyday activities. The type of network used for different activities often varies depending on the desired purpose. Professional networks such as LinkedIn are used to stay in touch with colleagues and coworkers. Personal social networks including the popular Facebook platform enable people to engage with their friends and to follow their news updates. News media and social network services such as Twitter allow people to follow short news updates (tweets) of celebrities and friends. Recently, another type of social network has become highly popular attracting millions of users: *online social collaboration networks*. These networks enable people to collectively work on projects. An example of such a social collaboration platform is Github [1]. GitHub was launched in 2008 and enables people to work on public (open source) or private projects. Indeed, open source development has a long history (e.g., see [2]) and dates back to the 1950s and 1960s when IBM released software sources of its operating systems and other programs [3].

A novel aspect of recent online social collaboration platforms on the World Wide Web such as GitHub is that they provide improved support for social networking features such as followers/followings or news feeds based on the users' social network. These online collaboration platforms enable users to discover interesting projects and repositories more quickly and let people collaborate almost

* Tel.: +43 6802049749.
E-mail address: daniel.schall@siemens.com

0950-5849/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.infsof.2013.12.003

instantaneously. An intriguing hypothesis was postulated by [4] arguing that GitHub will be the next big social network driven by what people do instead of who people know. In professional networks such as LinkedIn people are mainly connected because they know each other from, for example, past work experience.

In networks such as LinkedIn or Facebook friendship is represented as reciprocated links in an undirected graph. Services such as Twitter and recently GitHub are based on a directed network approach. A directed network approach allows users to follow other users based on their interest without requiring them to reciprocate the relationship. In traditional social networks, some users may be followed by many people without following many peers themselves ('stars' or 'celebrities'). Is this also the case for online social collaboration networks such as GitHub? People in GitHub are mostly followed because they work on interesting projects. Thus, this difference between conventional social networks and online social collaboration networks requires a novel 'who to follow' recommendation approach. One important aspect in knowledge-intensive disciplines such as software engineering is to promote the effective dissemination of knowledge [5]. The authors in [6] found that formal routines should be supplemented by collaborative and social processes to promote awareness and learning. In our opinion, follower networks provide excellent means to address the need for effective dissemination of knowledge through informal relations and information interest. Following the right person is essential to get information updates from the community leaders and 'gurus'. 'Who to follow' recommendations aim to solve the problem of selecting the right person to follow.

Follower networks, information flows through re-tweets, and 'who to follow' recommendations have been analyzed in great detail for platform such as Twitter [7–9] or in enterprise social media networks such as WaterCooler [10]. To our best knowledge, there is no existing work that proposed context-sensitive following recommendations in online development networks.

In this research we present the following key contributions:

- **Who to follow recommendation.** Here we propose a method and algorithm for 'who to follow' recommendations. 'Who to follow' recommendations can be based on *behavior*, *network*, or *similarity*. Our approach is based on network analysis techniques. User relevance with respect to following recommendations is based on a novel authority metric. Authority in this work means being an expert or guru in a specific area (e.g., expert/guru in 'javascript' programming). The approach is specifically targeted at online software development communities but may be applied to other types of collaboration networks as well.
- **Social network metrics and evaluation.** We analyze our approach by using social network (follower network) and activity data from GitHub. We introduce the used dataset and calculate various metrics such as reciprocity to support our hypothesis that people in GitHub are mostly followed because they work on interesting projects. The proposed authority-based 'who to follow' recommendation approach is evaluated by using the GitHub dataset.
- **Social network resilience.** The main purpose of follower networks such as the feature in GitHub is to get information updates and to potentially forward information (analogous to tweets and re-tweets). To understand social network robustness and resilience, we gradually remove nodes from the network and analyze metrics such as average path length. How many hopes on average must be passed to disseminate an item of information? We use datasets of other popular social networks such as Facebook, Twitter and Google Plus and compare them with GitHub.

This paper is structured as follows: Section 2 discusses relevant related work. Section 3 introduces our 'who to follow' recommendation approach. Section 4 introduces social network metrics and our evaluation. The paper is concluded in Section 5 with an outlook on future research.

2. Related work

We structure related work into relevant topics including analysis of online development communities, social network analysis, and social resilience.

2.1. Online development communities

An online social network is a communication and collaboration medium that connects a large number of people. People within the social network stay together if their interaction dynamics leads to the emergent property that is called 'community'. Here we focus on online development communities consisting of people developing collaboratively open source projects. A topological analysis of the SourceForge community was presented in [11]. The focus of the work was on role detection of users and cluster analysis. In [12], metrics with regards to open versus private software development were analyzed with the focus on source code aspects. Measures to investigate the social–technical congruence in software development projects were established in [13]. The interplay between network metrics, software failures and software evolution was investigated in, for example [14–16]. Collaboration and influence on GitHub was analyzed in [17] with the central focus on visualization techniques. Interesting directions with regards to the analysis of GitHub were presented in [18]. The authors [18] showed evidence for social collaboration on GitHub and proposed algorithms addressing the team formation problem. Our authority-based recommendation approach can well be used to create expertise profiles that can be used to assist in the formation of expert teams [19,20]. From a technical point of view, the basic structure of the GitHub API and the event schema was described in [21].

In this work, we analyze the GitHub online development community but focus on follower/following recommendations.

2.2. Social network analysis

Social network analysis techniques offer a rich set of theories (e.g., social network theory, small world phenomenon, power-laws, self-organization, and graph theory) and tools to analyze the topological features and human behavior in online communities [7,22–27].

In many systems, including large-scale enterprises, mostly searchable directories or databases that include descriptions of the employees' knowledge and experience are used to locate experts. The problem with this approach is that social networks and also big companies are in a constant state of flux and change [28,29]. In large-scale online communities and dynamic organizations, it becomes infeasible to constantly review and update the profile information of often rapidly changing experience, skills, and rolls of experts. Specifically with respect to software engineering, the authors in [30] found that a person with the most modifications to the code may be an expert within a community, and that expertise depends on the area of the code that is being modified. Furthermore, the 'degree of knowledge model' [31], showed how that expertise decays with subsequent changes by other authors.

We apply well-grounded theories and algorithms to the analysis of large scale software development communities. Well known ranking algorithms to calculate importance in linked environments include the HITS algorithm [32] and PageRank [33]. Personalization

Please cite this article in press as: D. Schall, Who to follow recommendation in large-scale online development communities, Inform. Softw. Technol. (2013), http://dx.doi.org/10.1016/j.infsof.2013.12.003