Expert Systems With Applications 54 (2016) 265–283

Contents lists available at ScienceDirect

Expert Systems With Applications

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

A multi-user perspective for personalized email communities

Waqas Nawaz[a,b], Kifayat-Ullah Khan[a], Young-Koo Lee[a,*]

[a] Dept. of Computer Engineering, Kyung Hee University Seocheon-dong, Giheung-gu, Yongin-si, Gyeonggi-do, 446-701, Korea
[b] Institute of Information Systems, Innopolis University, Universitetskaya St. 1, Innopolis, Tatarstan Republic, 420500, Russia

ARTICLE INFO

Keywords:
Graph clustering
Community detection
Collaborative similarity
K-Medoid clustering
Entropy
Density
F-measure
Personalized
Multi-user
Email network
Social media
PI-Net

ABSTRACT

Email classification and prioritization expert systems have the potential to automatically group emails and users as communities based on their communication patterns, which is one of the most tedious tasks. The exchange of emails among users along with the time and content information determine the pattern of communication. The intelligent systems extract these patterns from an email corpus of single or all users and are limited to statistical analysis. However, the email information revealed in those methods is either constricted or widespread, i.e. single or all users respectively, which limits the usability of the resultant communities. In contrast to extreme views of the email information, we relax the aforementioned restrictions by considering a subset of all users as multi-user information in an incremental way to extend the personalization concept. Accordingly, we propose a multi-user personalized email community detection method to discover the groupings of email users based on their structural and semantic intimacy. We construct a social graph using multi-user personalized emails. Subsequently, the social graph is uniquely leveraged with expedient attributes, such as semantics, to identify user communities through collaborative similarity measure. The multi-user personalized communities, which are evaluated through different quality measures, enable the email systems to filter spam or malicious emails and suggest contacts while composing emails. The experimental results over two randomly selected users from email network, as constrained information, unveil partial interaction among 80% email users with 14% search space reduction where we notice 25% improvement in the clustering coefficient.

1. Introduction

Email has become one of the most imperative asynchronous human communications. Many email systems,namely Gmail, Hotmail and Yahoo-mail, provide services to exchange information among users. Recent statistics show that 4087 million email accounts exist and 200 billion emails are exchanged worldwide each day (Radicati & Hoang, 2011). It is non-trivial for email systems to provide customized services to users based on the huge amounts of electronic data (Dabbish & Kraut, 2006; Wattenberg, Rohall, Gruen, & Kerr, 2005). Therefore, personalized information is utilized for this purpose, e.g. Gmail introduces a generic filter to group emails automatically into five categories (Gaikar, 2013). Our focus is to assist these systems for providing personalized services such as email strainer, malicious email identification, email group prediction, contact suggestions while composing emails, and guilt by association.

The analysis of email data[1] is different from other off-line social network analysis (Johnson, Kovcs, & Vicsek, 2012; Liu, Qu, & Wang, 2015; Qu, Liu, Yang, & Jensen, 2014). Community detection (Fortunato, 2010), as a major topic in network analysis, has received a great deal of attention with the knowledge of entire email network (Liu, Wang, Wang, Yao, & Liu, 2007; Moradi, Olovsson, & Tsigas, 2012). Discovering inherent community structures can help us understand the email network more deeply and reveal interesting properties shared by the members. People belonging to the same community are expected to have similar communication behaviors. Therefore, the identified communities can be used for classifying emails, discovery of prominent users, and highlighting abnormal activities inside the network (Johansen, Rowell, Butler, & Mcdaniel, 2007; Lin, 2010; Liu, Wang, Liu, & Zhang, 2009; Martin, Sewani, Nelson, Chen, & Joseph, 2005; Nagwani & Bhansali, 2010; Shetty & Adibi, 2005; Tan, Zhu, Qu, & Liu, 2014; Timofieiev, Snasel, & Dvorsky, 2008; Wilson & Banzhaf, 2009; Yang, Luo, Liu, Yin, & Cao, 2010; Yelupula & Ramaswamy, 2008; Yoo, Yang, Lin, & Moon, 2009).

[*] Corresponding author. Tel.: +82 31 201 3732; fax: +82 31 202 3706.
E-mail addresses: w.nawaz@innopolis.ru (W. Nawaz), kualizai@khu.ac.kr (K.-U. Khan), yklee@khu.ac.kr, w.nawaz@innopolis.ru, wicky786@gmail.com (Y.-K. Lee).

[1] http://www.email-marketing-reports.com/metrics/email-statistics.htm.

http://dx.doi.org/10.1016/j.eswa.2016.01.046
0957-4174/© 2016 Elsevier Ltd. All rights reserved.

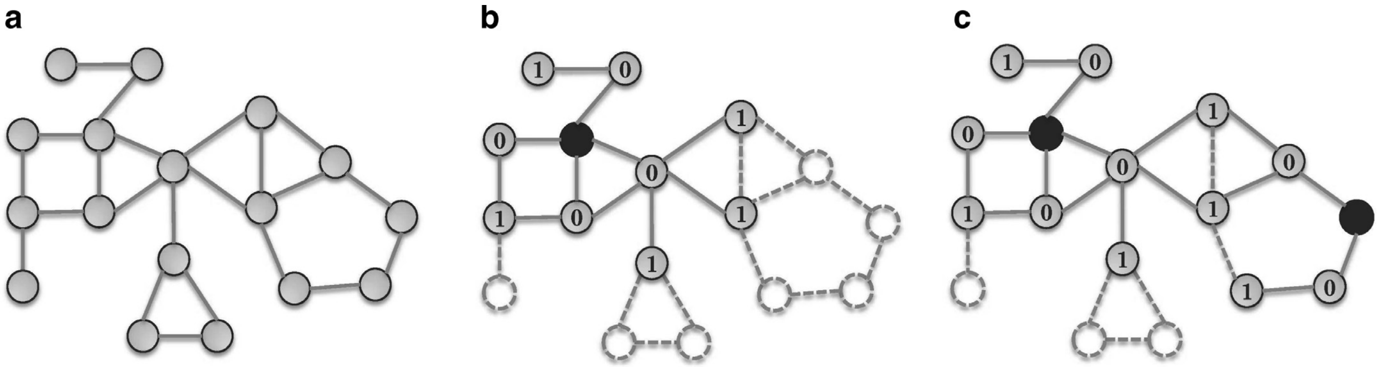

Fig. 1. Network information perimeter: (a) global, (b) personalized (or local), (c) multi-user personalized.

It is essential to utilize structural and contextual network information to determine the communication behavior of the users. In structural approaches, the relationships among individuals based on their interactions are analyzed in an email network, which is a kind of social network. In order to analyze the structure of email networks, several Social Network Analysis (SNA) techniques are adopted, such as node centrality (Freeman, 1978), cluster analysis (Clauset, Newman, & Moore, 2004; Papadopoulos, Kompatsiaris, Vakali, & Spyridonos, 2012), and topological structure (Ahuja, 2000; Zhu, Zhang, & Qu, 2013). This kind of analysis reflects either similar neighborhood structures for email communications, e.g., frequent email exchanges with shared neighbors, or communication intensity. The unification of structural and semantic information is also achieved for community analysis (Liu et al., 2007; Zhao et al., 2012).

Recently, Boykin and Roychowdhury and Yoo et al. claim that a global social network may include noisy features and deemphasize personalization in the inductive learning of important features through the network. It may also affect the user's own communication behavior or pattern. Consequently, the concepts of localization (Hu & Lau, 2012) or personalization (Boykin & Roychowdhury, 2004; Yoo et al., 2009) are introduced in literature. Inherently, both local and personalized approaches utilize the personal information owned by a particular user, e.g., all emails from single account. In other words, the personalization concept limits the topological information of the email network up to two-hops from a reference user to identify the communities. Fig. 1, shows the categorical views of the network information where each node and link represent a user and email exchange respectively. The global view refers to entire network information as depicted in Fig. 1(a). The other views differ by the reference user and its neighborhood. Personalized (or local) view follows the omni-guided equal bounds, i.e., limited to two hops in all directions. For example, in personalized view the reference user is represented by black (solid) node. All the nodes that are two hops away from a reference user are considered as known information.

However, communities are identified using extreme views of the email network, from global to local or personalized, which oscillate between efficiency and effectiveness. Moreover, it is almost impossible to approximate the entire email network community structure under the traditional concept of personalization. In order to achieve better communities at reasonable cost we are relaxing the personalized view of the email network using multi-user information. For instance, emails from more than one account can lead us beyond two-hop view of the network as shown in Fig. 1(c). Moreover, there is marginal probability for all the sender/receiver email IDs of each account to be mutually exclusive. It is an interesting and challenging issue to analyze the community structure using multi-user (or multi-account) information under the constraint of privacy. For example, users by exchanging frequent emails through different accounts of the reference user (i.e., owner of the account) with similar behavior are expected to be in the same community. So it requires an effective strategy to explore the multi-account email data for valuable insights in terms of communities.

In this paper, we present a personalized community detection method over multi-user email network, which is solely based on emails extracted from multiple email accounts. Personal emails of each user describe social activities that are transformed to an undirected weighted graph for structural and semantic analysis. Each user, i.e., either sender or receiver, is represented by a node and an edge reflects shared emails, where frequency is associated as an edge weight. The first phase extracts the communication patterns of interest (*CPI*) using multi-user emails as informative features to describe the communication behavior of each user. Subsequently, the second phase detects user communities via an intra-graph clustering method by contemplating structural and semantic aspects together. The semantic resemblance among individuals is achieved through their *CPI*s. We validate the effectiveness of the proposed technique on real email dataset in terms of various performance measures, i.e., density, entropy, and f-score. We also provide comparative analysis on community dynamics in terms of single and multi-user personalized information. Significant contributions of this work, in comparison to existing studies, are summarized as follows:

- **Multi-user personalized communities:** We introduce the notion of multi-user personalization under the constraint of privacy and unavailability of an entire email corpus. We uniquely construct an undirected, weighted, and multi-attributed graph using emails meta-data from more than one accounts. This enriched representation of email data enables the generic graph clustering approach to partition the vertices (users) effectively. These user groupings make the email system intelligent enough to filter and group the emails automatically based on similar communication patterns.
- **Community evolution:** This study uniquely investigates the dynamics of personalized communities in terms of network properties including density, avg. no. of neighbors, network centralization, clustering coefficient, and no. of vertices along with the visual analysis. The community changes is one of the important indicators to detect fraudulent account in email systems.
- **Personalization towards approximation:** Analyzing the community structure of entire email network of millions of users is computation intensive task for email systems. Multi-user personalization concept provides a mechanism to approximate the entire network community structure with partial information, i.e., a subset of email accounts.