A Survey of Mining Software Repositories in Social Network

Yuexiao Teng*
East China University of Science and Technology, Shanghai, China.

*Corresponding author. E-mail: roger.yxt@aliyun.com

Manuscript submitted January 10, 2020; accepted March 18, 2020.
doi: 10.17706/jsw.15.2.62-67

Abstract: Mining Software Repositories can serve a variety of functions e.g. predicting future software Engineering changes, studying code coverage. Social networks bring people on various locations together and mining social networks usually aims to find people's behavior patterns. However, there are few comprehensive surveys on the intersection of mining software repositories and mining social networks due to the little emphasis on this intersection and the recent rapid popularity of social networks e.g. GitHub. In order to fill this existing gap, in the paper, an original literature survey has been conducted on 5 different kinds of social networks regarding mining software repositories since 2000. To author's best knowledge, it is the first time to conduct such a survey on mining software repositories in social network.

Key words: Data mining, mining software repositories (MSR), social network, survey.

1. Introduction

Source code repositories, bug tracking systems, communication archives etc. are helpful in developing, testing as well as maintaining software products. Mining Software Repositories, seen as Software Knowledge Discovery in software databases is the automated or manual extraction of patterns which represent software knowledge stored in large software repositories. Mining Software Repositories is more than just kind of data mining whose sources come from software [1]. It is due to the fact that it requires domain-specific knowledge in software engineering. Currently, a number of Mining Software Repositories techniques have been used to parse, extract software data. For example, CodeWorker, an open-source software, can be trained to parse almost any language and provides distinct methods for creating parsers [2]. Another example is X-Diff, adopting a fairly efficient change detection algorithm, is a good tool for finding the difference between two XML software source files in a practical setting.

Many developers, testers and managers work on social networks on a daily basis. For example, modern IT companies use Email as a communication tool in the workplace to discuss development tasks, bug reproduction and customers' requirements and so on. As a result, this offers opportunities to mine Email archives to find useful patterns and subsequently predict the future. Additionally, with the further rise of the Internet, there are increasingly people who communicate by new social network such as Twitter, Facebook and WeChat. Information Technology (IT) practitioners are no exceptions, which means that these social networks can be mining targets for software engineering purposes.

Currently, there are a number of surveys on mining software repositories e.g. A Survey on Mining Software Repositories [1]. However, few well-rounded surveys have hitherto been carried out on mining social networks e.g. developers' blogs, as software repositories. There are two major reasons for this academic gap. First of all, many researchers working on mining software repositories lay little emphasis on mining social network software repositories. It is because they might consider it is less important than traditional software repositories such as source code or runtime logs. In addition, some social networks used by software team members are just recently of popularity e.g. GitHub, Stackoverflow. There have not been sufficient time to carry out the related research work so far. Therefore, in this paper I would like to literally survey the status of mining social networks as software repositories since 2000.

The contribution of the paper is that it is the first work, to author's best knowledge, to comprehensively conduct a literature survey of mining software repositories in the context of social network e.g. online community, blog.

Paper Organization: the rest of the paper is structured as below. Section 2 presents a brief overview of mining software repositories in social network. Section 3 surveys the intersection of mining software repositories and mining social networks. Section 4 draws a conclusion.

2. Mining Software Repositories in Social Network

In this section, an overview of Mining Software Repositories in social network would be presented and then the process of it would be illustrated in Fig. 1.

The description of Mining Software Repositories is a field which analyzes the rich available data in software repositories to uncover interesting and actionable information about software systems and projects [3]. Mining Software Repositories is defined as the process of automatically discovering useful information in large data repositories [4]. Mining social network usually focuses on people's group behavior from a sociological aspect, dating back to at least 70 years ago. The relationship between Mining Software Repositories and mining social networks is that Mining Social Networks and Mining Software Repositories have overlapped and could be considered that mining social networks is within the scope of data sources from software repositories. Thus, mining software repositories in social network focuses on mining social networks as software repositories. Therefore, Mining Software Repositories in social network needs not only software engineering specialized knowledge but also social network domain knowledge, possibly involving software testing and software development regarding social network.

The process of Mining Software Repositories in social network is shown in the Figure 1, which is similar to the process of normal mining software repositories. The noticeable difference is that mining targets are social networks as software repositories. Generally, there are three major steps.

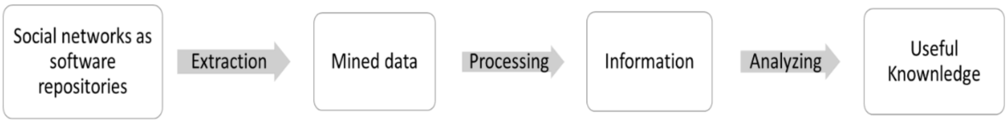

Fig. 1. An overview of mining software repositories process in social network.

3. The Survey on Mining Software Repositories in the Context of Social Network

In this section, I survey the literature on mining software repositories in social network. According to Developer Social Networks in Software Engineering: Construction, Analysis, and Applications [5], ZHANG WeiQiang et al. categorised developer social networks into Project Participation, Version Control System, Email Archives, Bug Tracking System and other. Thus, in the following part I conducted the literature survey on Email archives, online communities, blogs, bug tracking systems and version control systems. Additionally, I would mainly analyse social networks regarding mining software repositories from its research motivation, methodology, experimental results and qualified contributions in a reversely chronological way.

3.1 Mining Email Archives as Software Repositories

Since we live in an information age, working people make use of emails almost every day. Software practitioners likewise communicate by email, forming social networks in the work place. It is naturally believed to be a source of data for mining.

Liguo Yu *et al*. [6] investigated the interactions of distributed open-source software developers, leveraging clustering data mining techniques on email archives to study a representation of developer social network. They conducted empirical studies on two open-source projects and three related social networks. Their research focused on the dynamics feature of social networks. They found that the three open-source developer networks such as Linux or K Desktop Environment(KDE) mailing list evolved over a time period with some particular patterns. Though they could not derive a more general conclusion, it was very original in term of mining email archives to study the dynamics of developer social networks.

Christian Bird *et al*. [7] mined historical data on open-source software projects to examine the relationships between the email archives and software development related issues. They used PostgreSQL servers as their experimental targets by mining data from mailing list archives. In addition, they analysed the data and found that the relationships between the level of email activity and other software development activities. Although their study was modestly limited on data gathering, based on the experimental preliminary data, their unique findings undoubtedly contributed to mining social networks as software repositories research.

Prior to the research, Rakesh Agrawal *et al*. [8] carried out a fairly close study on mining newsgroups but not rigorously email archives to study people social behaviour using newsgroups. One interesting finding was that people more frequently respond to a message when they disagree than when they agree. Achieving effective classifying people posting to newsgroups, they analysed the graph structure of email responses.

3.2 Mining Online Communities as Software Repositories

In the recent years, online communities e.g. GitHub which basically functions as an online technical community and provides developers with online code repositories where they can create issues, commit code and communicate the latest techniques, Stackoverflow, which is another online community helping developers to solve technical issues, are increasingly popular among technicians and researchers in the information technology industry. Meanwhile, there are some research efforts made to mine online communities as software repositories. Since it is common that development teams store their project code on GitHub or ask some technical questions on Stackoverflow, the data on online communities is rich.

Asher Trockman *et al*. [9] studied crypto currencies by observing development efforts and software progresses, mining GitHub as software repositories. They patiently and determinedly spent a year of daily observations on more than 200 open source crypto currencies and concluded that there is no sufficient evidence for a relationship between market capitalization and daily stars, forks and other activities. Mining GitHub and focusing crypto currencies are novel regarding mining social networks in software repositories.

Another research group Saikat Mondal *et al*. [10] has done the similar work, but they conducted their exploratory study on Stackoverflow rather than GitHub. The motivation of the research was simply that sometimes submitted code segments bring difficulties in reproducing reported issues. Their study methodology was manual and it took the team 200 man hours to achieve research outputs. Based on the results that 68% of issues can be reproduced after code modifications, they suggested that the reproducibility could be a new metric of question issues. Although they did experimental work manually not automatically by programming, mining Stackoverflow with eyes set on the reproducibility feature of question issues is fairly original.

Before the above research in this sub section, Eirini Kalliamvakou *et al*. [11] did empirical studies on GitHub to describe the data quality and properties from this online community. The surveys and interviews were primarily adopted as their research methods. The results showed despite the rich data, over 10 million git code bases, on GitHub, mining GitHub should be careful because it could bring about potential threats. It is because of some factors such as high proportion of personal and inactive projects. From a balanced view, the research team successfully analysed the opportunities and risks of mining GitHub.

3.3 Mining Blogs as Software Repositories

It is popular that people use blogs or microblogs to document their life, while many software practitioners likewise put their technical thoughts, working notes and learning materials on their blogs e.g. CSDN, named Chinese Software Developer Network, which aims to offer Chinese programmers a platform to communicate about Information Technology techniques and skills, DZone, a famous developer community, which has more than 1 million members or microblogs e.g. Twitter, one of the largest social media. Therefore, naturally there is abundant data which could be mined for specific engineering and research purposes.

Dennis Pagano *et al*. [12] reported an exploratory study on how developers write blogs to understand how developers take advantage of social media in contrast with traditional development tools e.g. SVN. The research method was that they observed the usage of the blogs and analysed the content of the blogs in 4 large open-source communities. The research results could be helpful in understanding the role of blogs in software development process. Since there are mere studies on mining blogs as software repositories, the work was quite novel. However, the study only examined the value of blogs regarding software engineering.

Yuan Tian *et al*. [13] aimed to assist software development efforts by mining the content of microblogs from Twitter. The team investigated the popularity and diffusion of microblogs and found that there are a lot of information e.g. job openings, questions in microblogs. Additionally, they concluded that microblogs are more largely diffused in Twitter. Compared with other studies on blogs, it was meaningful to focus on microblogs by mining software repositories, despite the only one case study on Twitter.

3.4 Mining Bug Tracking Systems as Software Repositories

Bug tracking systems offer places where software engineers, project leaders or department managers could engage in software project activities just like social networks. As a result, it may be a good source of software engineering data for mining. Currently, there are two mainstream bug tracking systems e.g. Bugzilla, Jira[14].

Boyuan Chen *et al*. 15] conducted replication and empirical studies on 21 open-source Java projects in order to investigate the logging practices, compared with previous studies on open-source C++ projects. They mined data from Bugzilla and Jira and successfully found that a large portion of log updates are for improving the quality of logs. As the authors claimed, it was the first time to study the logging practices based on Java open-source projects. Therefore, their research targets originally differed from others.

=Aiming to fill the gaps in research regarding affect detection in software components, Marco Ortu *et al*. [16] mined issue tracking systems by manually labelling 2,000 bugs comments and 4,000 groups of words posted by software developers for analysing sentiments purposes and studying the relationship between affects and software development. Although the research output was minimal, the team firstly focused on the emotional side of mining software repositories.

Lucas D. Panjer [17] and Ahmed Lamkanfi *el al*. [18], 2 different research teams, mined issue tracking systems to find some patterns, serving as predicting the future e.g. bug lifetimes. The contributions of the research are the goals of mining bug tracking systems, i.e. prediction, which could be of usefulness in software development management.

3.5 Mining Version Control Systems as Software Repositories

Version control systems play an important role in modern software development. There are a number of version control systems e.g. CVS, SVN as well as Git, more and more prevalent, which are mainstreamed by software engineers and development teams to manage source code versions and deliver good quality software. Version control systems also can be considered as social networks due to the fact that users can interact with each other through commit logs, which could be mined like software repositories. Lopez-Fernandez et al. firstly used Version Control System-Developer Social Networks VCS-DSNs in 2004 [19].

Giampiero Di Paolo *et al*. [20] studied how emotions influence software developers' performance. The research method was mining readily available data from GitHub, which was used as a kind of version control systems, to analyze software developers' outputs. Though the research efforts were mere, the contribution of the paper was originally studying the correlation between developers' achievements and their emotions.

Motivated by understanding the social networks of developers, Alexander C. MacLean *et al.* [21] mined Apache Software Foundation (ASF) Subversion repository commits logs from 2010 to 2011. Based on the mined datasets, they presented a graph representing developers' commit actions. As they claimed in the paper, the work along with the research tools they have developed could facilitate further discussion on collaborative behaviour among developers.

Thomas Zimmermann *et al*. [22] investigated parallel development behaviour by means of mining CVS records from 4 large open-source projects. The research questions e.g. What is the degree of parallel development? How frequently do conflicts occur during updates and how are they resolved? raised in the beginning of the paper have been answered properly in the following. Compared with Alexander C. MacLean *et al*. [21] mining ASF SVN commits data to study developers' collaborative behaviour as mentioned previously, it did fill the research gap by mining CVS activity data to study developers' parallel behaviour, though they did not derive a more common conclusion.

4. Conclusions

In this paper, I surveyed the status of mining software repositories in terms of mining social networks. It is because, to author's best knowledge, there have been few surveys on the combination of mining software repositories and mining social networks. Another motivation is that social networks e.g. blogs, online communities have gained a lot of popularity these days. Thus, a survey related to mining social networks can facilitate further research. To sum up, since 2000, there have been some research efforts made to mine social networks as software repositories primarily using the empirical studies, some of which were exploratory.

References

[1] Jung, W., Lee, E., & Wu, C. (2012). A survey on mining software repositories. *IEICE Trans. Inf. & Syst.*, 1384–1406.

[2] CodeWorker – Parsing tool and a source code generator. Retrieved from: http://codeworker.org/

[3] MSR 2011: 8th Working Conference on Mining Software Repositories. Retrieved from: http://2011.msrconf.org/

[4] Introduction to data mining: Pearson new international edition. Retrieved from: https://www.pearson.ch/HigherEducation/ComputerScience/DatabaseSystems/EAN/9781292026152/Introduction-to-Data-Mining-Pearson-New-International-Edition

[5] Zhang, W., Nie, L., Jiang, H., Chen, Z., & Liu, J. (2014). Developer social networks in software engineering: construction, analysis, and applications. *Sci. China Inf. Sci.*, *57(12)*, 1–23.

[6] Yu, L., & Ramaswamy, S., & Zhang, C. Mining email archives and simulating the dynamics of open-source project developer networks,.

[7] Bird, C., Gourley, A., Devanbu, P., Gertz, M., & Swaminathan, A. (2006). Mining email social networks in

postgres. *Proceedings of the 2006 International Workshop on Mining Software Repositorie*.

[8] Agrawal, R., Rajagopalan, S., Rajagopalan, S., Srikant, R., & Xu, Y. (2003). Mining newsgroups using networks arising from social behavior. *Proceedings of the 12th International Conference on World Wide Web*.

[9] Trockman, A., Tonder, R. V., & Vasilescu, B. (2019). Striking gold in software repositories?: An econometric study of cryptocurrencies on github. *Proceedings of the 16th International Conference on Mining Software Repositories*.

[10] Mondal, S., Rahman, M. M., & Roy, C. K. (2019). Can issues reported at stack overflow questions be reproduced?: An exploratory study. *Proceedings of the 16th International Conference on Mining Software Repositories, Piscataway*.

[11] Kalliamvakou, E., Gousios, G., Blincoe, K., Singer, L., German, D. M., & Damian, D. (2016). An in-depth study of the promises and perils of mining GitHub. *Empir Software Eng*, *21(5)*, 2035–2071.

[12] Pagano, D., & Maalej, W. (2011). How do developers blog?: An exploratory study. *Proceedings of the 8th Working Conference on Mining Software Repositorie*.

[13] Tian, Y., Achananuparp, P., Lubis, I. N., Lo, D., & Lim, E.-P. (2012). What does software engineering community microblog about? *Proceedings of the 2012 9th IEEE Working Conference on Mining Software Repositories (MSR)*.

[14] Atlassian. Jira | issue & project tracking software. Atlassian. Retrieved from: https://www.atlassian.com/software/jira

[15] Chen, B., & Ming, J. Z. (2017). Characterizing logging practices in Java-based open source software projects — A replication study in apache software foundation. *Empirical Software Engineering*.

[16] Ortu, M. *et al*., (2016). The emotional side of software developers in JIRA. *Proceedings of the 2016 IEEE/ACM 13th Working Conference on Mining Software Repositories*.

[17] Predicting eclipse bug lifetimes. (2019). *Proceedings of the Fourth International Workshop on Mining Software Repositories*.

[18] Lamkanfi, A., & Demeyer, S. (2013). Predicting reassignments of bug reports — An exploratory investigation. *Proceedings of the 2013 17th European Conference on Software Maintenance and Reengineering*.

[19] Applying social network analysis to the information in CVS repositories. 101–105.

[20] Paolo, G. D., Malavolta, I., & Muccini, H. (2014). How do you feel today? buggy! *Proceedings of the 2014 40th EUROMICRO Conference on Software Engineering and Advanced Applications*

[21] Apache commits: social network dataset. *Proceedings of the 10th Working Conference on Mining Software Repositories*.

[22] Zimmermann, T. (2007). Mining workspace updates in CVS. *Proceedings of the Fourth International Workshop on Mining Software Repositories*.

Yuexiao Teng received his computer science B.S. degree in 2009 and computer application technology M.S. degree in 2012 both from East China University of Science and Technology, Shanghai, China. He has several years working experience in information technology industry. Currently, his research interests are software engineering, software testing and data mining etc.