ISSN 2186-7437

NII Shonan Meeting Report

No. 2017-2

Mining Software Repositories: Accomplishments, Challenges and Future Trends

Emad Shihab
Akinori Ihara

March 6–10, 2017

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-Ku, Tokyo, Japan

Mining Software Repositories: Accomplishments, Challenges and Future Trends

Organizers:
Emad Shihab (Concordia University, Canada)
Akinori Ihara (Nara Institute of Science and Technology, Japan)

March 6–10, 2017

The field of Mining Software Repositories (MSR) has been steadily growing over the past decade. Since its first workshop in 2004, the MSR community has steadily grown to be one of (if not) the biggest co-located events with the International Conference on Software Engineering, the flagship conference in Software Engineering.

The main goal of the MSR community is to leverage development data, often stored in software repositories. Examples of these repositories are: source control repositories, which store source code changes, defect tracking repositories, which store software defect reports and communication repositories, which store developer communications such as emails. These repositories contain a wealth of information that is available for most software projects. The MSR community has proposed techniques to effectively mine repository data, leverage such data to improve requirement, quality and traceability of software project and empirically study the impact of several development phenomena.

However, now MSR is at a critical point where many accomplishments have been made, but many challenges remain due to the changing landscape of software engineering. For example, the availability of data was one of the biggest challenges for MSR in the past, whereas now the availability of too much data is causing challenges. Recent technology advancements in complementary areas such as machine learning and artificial intelligence has enabled MSR researchers to develop more accurate techniques, however usability remains as an open challenge. The widespread use of mobile devices has lead to more mobile-related software, often known as mobile apps, is also a new trend that the MSR community has recently started to target.

In this school, we invited 4 pioneers in MSR research field as a speaker and 25 up-and-coming researchers as a student. Given the recent accomplishments, these speakers taught and train future generations about the successes and future challenges of MSR. The students can learn from leaders in the MSR field and discuss potential solutions for such challenges.

Given the recent accomplishments, challenges and trends in MSR, we plan to organize a NSS to teach and train future generations about the successes and future challenges of MSR. The NSS will serve as a forum where students can learn from leaders in the MSR field and discuss potential solutions for such challenges. Through both hands-on and lecture sessions in this school, the lectures were given by a mixed set of well-established and emerging leaders in

the MSR field. We expect this school was given ample knowledge and practical approached by lectures and the students will show up exciting directions for improving MSR research in software engineering field to the next generation.
School Homepage: `http://nii-msrschool2017.se-naist.jp/`

Meeting Schedule

- Mar. 5, Sunday Evening
  - Welcome Reception
- Mar. 6, Monday Morning
  - Lecture1: "Mining Software Repositories: Accomplishments and Challenges", Ahmed E. Hassan
- Mar. 6, Monday Afternoon
  - Hands-on for Lecture 1, Ahmed E. Hassan
  - Panel 1: "Publishing in MSR", Ahmed E. Hassan, Daniel Morales German, Shane McIntosh, Alberto Bacchelli
- Mar. 7, Tuesday Morning
  - Lecture 2: "When history matters: using Software Repositories to address Source Code Provenance", Daniel Morales German
- Mar. 7, Tuesday Afternoon
  - Hands-on for Lecture 2, Daniel Morales German
  - Panel 2: "Collaboration Setting", Ahmed E. Hassan, Daniel Morales German, Shane McIntosh, Alberto Bacchelli
- Mar. 8, Wednesday Morning
  - Poster Presentation for Students
- Mar. 8, Wednesday Afternoon
  - Excursion and Dinner in Kamakura
- Mar. 9, Thursday Morning
  - Lecture 3: "Building on an unsound foundation: How release pipelines can impact our predictive models",
- Mar. 9, Thursday Afternoon
  - Hands-on for Lecture 3, Shane McIntosh
  - Panel 3: "Academic Career", Shane McIntosh, Alberto Bacchelli, Cor-Paul Bezemer, Akinori Ihara, Emad Shihab
- Mar. 10, Friday Morning
  - Lecture 4: "Supporting the human aspects of software engineering", Alberto Bacchelli
  - Hands-on for Lecture 4, Alberto Bacchelli

Overview of Talks

Mining Software Repositories: Accomplishments and Challenges

Ahmed E. Hassan, Queen's University

Software engineering data (such as code bases, execution traces, historical code changes, mailing lists, and bug databases) contains a wealth of information about a project?s status, progress, and evolution. Using well established data mining techniques, practitioners and researchers can explore the potential of this valuable data in order to better manage their projects and to produce higher quality software systems that are delivered on time and within budget. This lecture will present the latest research in mining Software Engineering (SE) data, discusses challenges associated with mining SE data, highlight SE data mining success stories, and outlines future research directions. Attendees will acquire the knowledge and skills needed to perform research or conduct practice in the field and to integrate data mining techniques in their own research or practice. A hands-on illustration of the commonly used analysis tools such as R and WEKA will also be provided.

When history matters: using Software Repositories to address Source Code Provenance

Daniel Morales German, Victoria University

Who owns the copyright a project? The answer to this question determines who can license the code (either commercially or open source). It is also an important question during business acquisitions. Unfortunately, this is not always an easy question to answer. On one hand, copying code is easy and there is lack of traceability in tools—such as editors, and version control systems - of such copying and its source. On the other hand, software development, specially in open source, is increasingly becoming a team effort. In the absence of contributor copyright assignments (that transfer the ownership of a contribution to the project) the ownership of the source code becomes difficult to asses. Even those who reuse open source need to be concerned that the software they are reusing is properly licensed.In order to answer the question of ownership of copyright one needs to first answer the question: "what is the provenance of this code?". In this lecture I will describe the challenges of provenance discovery. These challenges include: the discovery of reliable corpora, the use of Bertillonage and clone detection to identify copied code, the analysis of the history of development to asses who are copyright authors of a system. I will also describe the challenges that copyright law impose on legally defining how software modifications contribute (or not) to the overall copyright of a system. Finally, I will overview the research we have performed during the last years on provenance discovery, and how we have used software repositories to do it.

Building on an unsound foundation: How release pipelines can impact our predictive models

Shane McIntosh, McGill University

Mining Software Repositories (MSR) researchers use complex statistical regression models or machine learning techniques to understand software engineering phenomena. We apply MSR techniques to analyze historical data that is stored in software repositories. As the MSR field has matured, and MSR techniques have become more robust, the size of our studied datasets (in terms of number of projects) have grown. While this growth attempts to tackle natural external validity concerns, they increase internal validity risks. In this lecture, I will discuss the importance of understanding the release pipeline of our studied projects. I will elaborate on how naive treatment of files, releases, and branches can lead to noise and biases that threaten the validity of MSR analyses. Furthermore, I will provide a framework for how release pipeline biases can be addressed. A hands-on component showing how to extract and leverage release data from repositories will also be provided.

Supporting the human aspects of software engineering

Alberto Bacchelli, Delft University of Technology

Abstract: Software development leads to the creation of large amounts of data, such as source code changes, defects, and test executions. Software Analytics aims at uncovering patterns and actionable insights from this data to support the human aspects of software development and maintenance. Selecting the right data is key to the success of Software Analytics. Unstructured software data (e.g., emails, bug descriptions, and technical forum discussions) is a valuable form of data as it opens a unique view on human factors involved in a software project; yet it is hard to harness. In the first part of the talk, I introduce how automated techniques based on text search, machine learning, and island parsing can be used to mine this data and obtain actionable results. Data alone is not enough: It has to be analyzed to answer the right questions to tackle relevant developers' needs. In the second part of this lecture, I will introduce how qualitative research methods can be used to uncover developers' needs. Particularly, as an example, I describe how we uncovered motivations, real outcomes, and fundamental challenges of Modern Code Review, thus opening a very promising research line to be tackled with Software Analytics. I also present a few analyses that can be done to support modern code review with data.

List of Participants

- Alexander Schlie, Technische Universitytät Braunschweig , Netherlands
- Anand Ashok Sawant, Delft University of Technology, Netherlands
- Davide Wille, TU Braunschweig, Germany
- David Spadini, Delft University of Technology, Netherlands
- Gema Rodríguez Pérez, University Rey Juan Carlos, Spain
- Keitaro Nakasai, NAIST, Japan
- Kunihiro NODA, Tokyo Institute of Technology, Japan
- Luca Pascarella, Delft University of Technology, Netherlands
- Masanari Kondo, Kyoto Institute of Technology, Kyoto, Japan
- Marco di Biase, Delft University of Technology, Netherlands
- Natthawute Sae-Lim, Tokyo Institute of Technology, Japan
- Karim Md. Rejaul, NAIST, Japan
- Ștefan Stănciulescu, IT University of Copenhagen, Denmark
- Toshiki Hirao, NAIST, Japan
- Vladimir Kovalenko, Delft University of Technology, Netherlands
- Xin Yang, Osaka University, Japan
- Wu Yuhao, Osaka University, Japan
- Shade Ruangwan, NAIST, Japan
- Jirayus Jiarpakdee, NAIST, Japan
- Rabe Abdalkareem, Concordia University, Canada
- Cor-Paul Bezemer, Queen ' s University, Canada
- Ruiyin (Ray) Wen, McGill University, Canada
- Keheliya Gallaba, McGill University, Canada
- Ahmed E. Hassan, Queen ' s University, Canada
- Shane McIntosh, McGill University, Canada
- Alberto Bacchelli, Delft University of Technology, Netherlands
- Daniel Morales German, Victoria University, Canada