MSR’04: International Workshop on Mining Software Repositories

Theme and Goals of Workshop
Software repositories contain a wealth of valuable information for empirical studies in software engineering: source control systems store changes to the source code as development progresses, defect tracking systems follow the resolution of software defects, and archived communications between project personnel record rational for decisions throughout the life of a project. Such data is available for most software projects and represents a detailed and rich record of the historical development of a software system. Participants in multiple sites, often in multiple continents, develop software projects without ever meeting in person, as is the case in many large commercial and Open Source projects. This trend makes the use of tools to record all aspects of software project more critical.

Until recently, data from these repositories was used primarily for historical record supporting activities such as retrieving old versions of the source code or examining the status of a defect. Several studies have emerged that use this data to study various aspects of software development such as software design/architecture, development process, software reuse, and developer motivation. These studies have highlighted the value of collecting and analyzing these data. Yet each of these studies has built its own version of methodologies and tools to address the formidable challenge of utilizing such data to perform their empirical research. Several international efforts have identified the development of approaches to extract, share, and study this data as a research priority.

The goal of this one-day workshop is to bring together researchers, practitioners, and source control developers to consider methods to use the data stored in these software repositories to further understanding of software development practices.

Participants, Selection Process and Activities
We expect to have around 30 participants. The workshop will be one-day long, starting at 8:30am and going until around 6pm. Emphasis will be put on making the workshop interactive with many periods for discussion assigned throughout the schedule, by limiting the time allocated for presentations (15-20 mins including clarification questions). At the end of the day, we should have a wrap-up panel to discuss and set goals for further research in the area of mining software repositories.

A Program Committee will assist in the submission reviewing and selection process. Position papers should be at most 5 pages and may address issues including but not limited to the following:

- New approaches to analyze the data stored in software repositories to:
- Assist in program understanding and visualization
- Predict and gauge the reliability and quality of software systems
- Study the evolution of software systems
- Discover patterns of change and refactorings
- Understand the origins of code cloning and code design change
- Model software processes for development, defect repair, etc.
- Assist in project planning and resource allocation

- Case studies on extracting data from these repositories for large long lived projects
- Proposals for common exchange formats, meta-models, and infrastructure tools to ease the sharing of the extracted data and to enable reuse and repeatability of results throughout the research community
- Suggestions for particular large software repositories to be shared among the community for research evaluation and benchmarking purposes
- Approaches to integrate data between repositories and with other software project data such static or dynamic analysis data
- Requirements and guidelines for users and developers of source control systems to ease the analysis of the stored historical data

We are looking for papers that can serve as the basis for fruitful discussions. We will select papers so that a broad range of stakeholders from across the software engineering discipline will be represented in the workshop. The accepted papers will be posted on the workshop’s web site prior to the workshop and proceedings will be provided at the workshop.

**Expected Workshop Output**

We expect the presentations and discussions in this workshop will facilitate the definition of challenges, ideas and approaches to transform software repositories from static record keeping repositories to active repositories used by researchers to gain empirically based understanding of software development, and by software practitioners to predict and plan various aspects of their project.

**Program Committee**

The following are members of the Program Committee:

1. Harald Gall (Technical U. Of Vienna, Austria)
2. Les Gasser (U. of Illinois, Urbana-Champaign, USA)
3. Daniel German (U. of Victoria, Canada)
4. James Herbsleb (CMU, USA)
5. Katsuro Inoue (Osaka U., Japan)
6. Philip Johnson (U. of Hawaii, USA)
7. Dewayne Perry (U. of Texas, USA)
8. Andreas Zeller (Saarland U., Germany)

**Interested Participants**

Furthermore, we assembled a list of individuals with related work in the area that would be interested in the workshop. We contacted many of them and sought input about the workshop scope and details. We plan to contact the rest of the the list once the workshop is approved. We plan to advertise the workshop in several venues (IEEE
conferences/SEWorld) and through the help of the interested participants once it is approved. The names marked with a (*) have already expressed interest in participating to us:

1. Giulio Antoniol (U. of Sannio, Italy)
2. Brian Behlendorf (Collab.net)*
3. Jai Asundi (U. of Texas, USA)
4. Dirk Draheim (Freie Universitat Berlin, Germany)*
5. Michael Godfrey (U. of Waterloo, Canada) *
6. John Hudepohl (Nortel Networks, USA) *
7. Rainer Koschke (U. of Stuttgart, Germany)
8. Karim Lakhani (Harvard U., USA)
9. Tim Lethbridge (U. of Ottawa, Canada) *
10. Greg Madey (Notre Dame U., USA)
11. Stan Matwin (U. of Ottawa, Canada) *
12. Renee Miller (U. of Toronto, Canada)
13. Maurizio Morisio (Politecnico di Torino)
14. Gail Murphy (U., of British Colombia, Canada)
15. Lukasz Pekacki (Freie Universitat Berlin, Germany)*
16. Robert Sanduski (U. of California, Irvine, USA)*
17. Walt Scacchi (U. of California, Irvine, USA)*
18. Stephen Schach (Vanderbilt U., USA)
19. Harvey Siv (Lucent Technologies, USA)
20. Eleni Stroulia (U. of Alberta, Canada)*
21. Jim Whitehead (U. of California, Santa Cruz, USA)
22. Kenny Wong (U. of Alberta, Canada)*

Organizer's Backgrounds
The workshop is organized by:

Ahmed E. Hassan is a Ph.D. candidate at the University of Waterloo. He received his MMATH and BMATH degrees from the school of Computer Science at UW in 1999 and 2001. His research interests include mining source control data, visualization and migration of web applications, and software architecture and evolution. He has over six years of industry experience working at IBM's Almaden Research Lab, Nortel and Research In Motion (RIM). He holds several patents in the area of wireless communications and distributed systems. (See WWW http://plg.uwaterloo.ca/~aec hassa)

Richard C. Holt is a Professor at the University of Waterloo, where his research interests include visualizing software architecture. This work includes reverse engineering of legacy systems and repairing software architecture. His architectural visualizations have included Linux, Mozilla (Netscape), IBM's TOBEY code generator, and Apache. His previous work includes foundational work on deadlock, development of a number of compilers and compilation techniques, development of the first Unix clone, and authoring a dozen books on programming and operating systems. He is one of the designers of the Turing programming language. (See WWW http://plg.uwaterloo.ca/~holt)

Ric has co-organized WoSEF (Workshop on Standard Exchange Format held at ICSE 2000) - http://www.cs.toronto.edu/~simsuz/wosef/. In addition, he was the general chair for IWPC 2001 - http://www.swen.uwaterloo.ca/~iwpc2001/

Audris Mockus conducts research of complex dynamic systems. He designs data mining methods to summarize and augment the system evolution data, interactive visualization techniques to inspect, present, and control the systems, and statistical models and optimization techniques to understand the systems. Audris Mockus received B.S. and M.S. in Applied Mathematics from Moscow Institute of Physics and Technology in 1988. In 1991 he received M.S. and in 1994 he received Ph.D. in Statistics from Carnegie Mellon University. He works at Software Technology Research Department of Avaya Labs. Previously he worked at Software Production Research Department of Bell Labs. His CV is at http://m okus.us/amvita.pdf

Audris is the program chair for the 10th International Symposium on Software Metrics - http://swmetrics.org/