3rd Workshop on Mining Unstructured Data

Alberto Bacchelli*, Nicolas Bettenburg†, Latifa Guerrouj‡, and Sonia Haiduc§
*Delft University of Technology, The Netherlands and University of Lugano, Switzerland
†Software Analysis and Intelligence Lab - Queen's University - Kingston, ON, Canada
‡SOCCER Lab - Ecole Polytechnique de Montreal - Montreal, QC, Canada
§Department of Computer Science - Florida State University - Tallahassee, FL, USA
a.bacchelli@tudelft.nl, nicbet@cs.queensu.ca, latifa.guerrouj@polymtl.ca, shaiduc@cs.fsu.edu

***Abstract*—Software development knowledge resides in the source code and in a number of other artefacts produced during the development process. To extract such a knowledge, past software engineering research has extensively focused on mining the source code, *i.e.*, the final product of the development effort. Currently, we witness an emerging trend where researchers strive to exploit the information captured in artifacts such as emails and bug reports, free-form text requirements and specifications, comments and identifiers. Being often expressed in natural language, and not having a well-defined structure, the information stored in these artifacts is defined as unstructured data. Although research communities in Information Retrieval, Data Mining and Natural Language Processing have devised techniques to deal with unstructured data, these techniques are usually limited in scope (*i.e.*, designed for English language text found in newspaper articles) and intended for use in specific scenarios, thus failing to achieve their full potential in a software development context. The workshop on Mining Unstructured Data (MUD) aims to provide a common venue for researchers and practitioners across software engineering, information retrieval and data mining research domains, to share new approaches and emerging results in mining unstructured data.**

I. Motivation

To analyze, comprehend, and reverse engineering software projects and their software development processes, we rely on various sources of information. Bug reports, execution logs, mailing lists, code review reports, change logs, requirements documents, and the actual source code contain implicit developer knowledge about the project and past development efforts. Most of this knowledge is captured as unstructured information: Natural language text used to exchange information among people. Researchers in the Information Retrieval (IR), Data Mining (DM), and Natural Language Processing (NLP) fields have experimented with various techniques (such as Latent Dirichlet Allocation and hierarchical clustering) and ad-hoc approaches to enable the mining of unstructured data. However, these techniques were not designed to work with the complexities and peculiarities of unstructured software engineering data, thus are not readily applicable to the software engineering research domain.

The MUD 2013 workshop aims to tackle these challenges and make mining unstructured data clear, accessible, and applicable to the software engineering domain. We propose to achieve this via three parts of the workshop. First, we invite peers to give a written description of their experiences with mining unstructured data, in the form of short (4-page) papers to be presented at the workshop, by sharing the techniques they used, the challenges they faced, and the solutions that they found successful. Second, we encourage discussion and dissemination of the presented work in following extended group discussion sessions. Third, we organize a group discussion in the form of a panel, according to the "fishbowl" technique, to identify and discuss MUD topics that are most relevant to the workshop participants. By collecting available techniques, solutions, and challenges yet to be overcome, we aim to advance the state-of-the-art in mining unstructured data.

II. Topics

MUD 2013 aims to address the following topics :

1) Applications of unstructured data mining techniques to support software maintenance, software reverse engineering tasks (e.g., feature location, traceability), and for enhancing software quality;
2) Novel sources of unstructured data, such as phone records, screenshots, interviews, or wiki pages;
3) Usage of NLP, IR, and ML techniques for mining unstructured data;
4) Classification and dissemination of techniques for extracting unstructured data;
5) Identification of open challenges and proposed solutions;
6) Novel extractors for unstructured data and performance evaluation with respect to existing techniques;
7) Linking of unstructured and structured data;
8) Large-Scale mining of Unstructured Data in Big Data environments.

978-1-4799-2931-3/13/$31.00 © 2013 IEEE

III. Goals and Expected Results

The MUD 2013 workshop aims to provide an interactive venue for researchers and practitioners interested in Mining Unstructured Data. The intended outcomes of this workshop are to:

1) Facilitate knowledge-exchange in the field of mining unstructured data and practical applications of MUD techniques through presentations of short (4-page) paper submissions.
2) Establish connections between the various research communities that mine unstructured data, resulting in cross-fertilization of techniques and methodologies.
3) Put techniques and methodologies for mining unstructured data in a common framework, enabling researchers and practitioners to find the appropriate tools that meet their particular data mining needs.
4) Identify open problems and challenges for mining unstructured data, providing the basis for a roadmap of future research opportunities in mining unstructured data research.
5) Educate on, discuss, and advance the state-of-the-art in mining unstructured data.

IV. Format

We propose MUD 2013 as a half-day workshop, consisting of (1) an introductory presentation, (2) a keynote, (3) presentations of short papers as a base for semi-structured group discussions, and (4) a closing fishbowl-panel discussion. We invite researchers in the field to present techniques that they have adapted from NLP, IR, and ML to mine Unstructured Data, as well as case studies of use of these techniques in practical applications. We encourage workshop participants to share their experience in mining unstructured data with the community, including negative research results, *i.e.*, "what did not work". Invited experts will attend the discussion panel to give additional feedback to workshop presenters and moderate a discussion on the state-of-the-art of mining unstructured data. The introductory presentation places mining unstructured data in the context of reverse-engineering, whereas the keynote speaker provides an account of personal experiences and faced challenges. Invited workshop participants will provide attendees with ideas, comments and feedback. Presentations of workshop papers are allocated a 30 minute time slot, divided into 15 minute paper presentation, followed by an extended 15 minute discussion and dissemination of the work by workshop participants. We strongly believe that balancing a keynote and invited experts with extended group discussion is essential for meeting the MUD workshop's goal.

V. Organizers

Alberto Bacchelli is currently an Assistant Professor at Delft University of Technology. He received his PhD in 2013 at the University of Lugano, Switzerland, under the supervision of Prof. Michele Lanza in the REVEAL (Reverse Engineering, Visualization, Evolution Analysis Lab) research group. His research interests include empirical software engineering, mining software repositories, unstructured data mining, software quality, and development tools for software engineering. In addition to his service on various conference committees, he was the co-organizer of the MUD'12 workshop.

Nicolas Bettenburg is a PhD student at Queen's University (Canada) under the supervision of Dr. Ahmed E. Hassan. His research interests are in mining unstructured information from software repositories with a focus on relating developer communication and collaboration to software quality. In the past, he has co-organized various conference tracks and has been a co-organizer of the MUD workshop since 2010.

Latifa Guerrouj is a PhD Student at the department of computing and software engineering of Ecole Polytechnique of Montreal. Her research areas are program comprehension and software quality, in particular through the development of theories, approaches and tools that ease program understanding and enhance the quality of source code. Latifa is serving as a program committee member of WCRE'13 and MSR'13 - data track. She also co-organized MUD'12 collocated with WCRE'12.

Sonia Haiduc is currently an Assistant Professor at Florida State University. She received her PhD from Wayne State University in 2013. Her research interests are in software maintenance, specifically in applying Information Retrieval and Natural Language Processing techniques to support tasks such as concept and feature location, code summarization, etc. She has served on several program and organizing committees for conferences in the field, and is currently the publicity chair for SCAM'14 and social media chair for ICPC'14.