Workshop on Mining Unstructured Data (MUD)

. . . because "mining unstructured data is like fishing in muddy waters"!

Alberto Bacchelli*, Nicolas Bettenburg† and Latifa Guerrouj‡
*REVEAL @ Faculty of Informatics - University of Lugano - Lugano, Switzerland
†Software Analysis and Intelligence Lab - Queen's University - Kingston, ON, Canada
‡SOCCER Lab - Ecole Polytechnique de Montreal - Montreal, QC, Canada
Email: *alberto.bacchelli@usi.ch, †nicbet@cs.queensu.ca, ‡latifa.guerrouj@polymtl.ca

***Abstract*—Software developers have long been supported by a variety of tools, such as version control systems (*e.g.*, GIT), issue tracking systems (*e.g.*, BugZilla), and mailing list services (*e.g.*, Mailman). These tools accumulate a wide range of information that is recorded in the repositories these tools store their data in. This information is comprised of two significantly different types of data: *structured* and *unstructured* data. Structured data (*e.g.*, source code or execution traces) has a well-established structure and grammar, thus is straightforward to parse and use with computer machinery. Unstructured data (*e.g.*, documentation, discussions, comments, or customer support requests) consists of a mixture of natural language text, snippets of structured data, and noise. Mining unstructured data is very challenging since out-of-the box approaches adopted from related fields, such as Natural Language Processing and Information Retrieval, cannot be directly applied in software engineering.**

To tackle challenges faced when mining unstructured data and make the knowledge contained in unstructured data repositories accessible to both practitioners and researchers, we organize the 2^{nd} workshop on Mining Unstructured Data (MUD'12). The aim is to provide a unique interactive venue for discussing in-depth challenges, approaches, and applications and share experiences, and results on the topic of mining software unstructured data.

I. Motivation

Software development tools such as version control systems, issue trackers, and mailing list services produce and record a large amount of information stored in software data repositories. Researchers extract knowledge from this data by mining these repositories, both to empirically validate novel research ideas, and to support practitioners' day-to-day activities such as program comprehension, reverse engineering, or re-documentation tasks.

Many data sources consist of information that is well structured, because it comprises artifacts either written by humans for a machine (*e.g.*, source code, formal specifications and models) or generated by a machine for humans (*e.g.*, execution traces). The knowledge embedded in *structured data* can currently be extracted and modeled through well-established techniques.

Other software data sources archive data that is more unstructured, as it is produced by humans for humans: documents, such as emails, change comments, or bugs' reports, written in natural language and used to exchange information among people. The information stored in *unstructured data* is valuable for practitioners and researchers alike, as it encodes developer knowledge not to be found in other artifacts. Unstructured data, by its nature of lacking structure, is more challenging for mining than structured data.

To mine unstructured data, researchers have been experimenting with technologies adopted from related research fields. Techniques such as topic models from Information Retrieval (IR), hierarchical clustering from Data Mining (DM), or Natural Language Processing (NLP) have been proven limited, in a sense that they are often laboriously tailored to the intricacies of the underlying data and intended use cases. As a result, a plethora of hand-crafted techniques emerged and have been proposed to mine unstructured data. The ad-hoc nature and terse documentation of these techniques, however, hinder their use for other tasks: this variety makes it hard for researchers and practitioners to determine the appropriate technique(s) to deal with the problem at hand and ways to use the selected technique effectively.

To address these challenges, and make the unstructured data contained in software development repositories more accessible and transparent, we organize the 2^{nd} workshop on Mining Unstructured Data (MUD'12). MUD'12 is intended to provide a unique forum for researchers across multiple domains to advance the state-of-the-art in mining unstructured data, provide a common framework for sharing new experiences, present, compare, and evaluate new techniques in the area of mining unstructured data.

II. Topics

We seek contributions related to techniques and applications of mining unstructured data. Contributions would ideally, but are not limited to, address one of the following topics:

1) Tools and methods for extracting unstructured data from software development repositories.
2) Adaptations of NLP and IR approaches to the software engineering domain.
3) Lessons learned and pitfalls of working with unstructured data and potential solutions.
4) Analysis of unstructured data embedded within structured data (e.g., comments in source code).
5) Unstructured data linking and traceability.
6) Usage of unstructured data for software evolution tasks, such as Feature/Concept Location, evolution of unstruc-

1095-1350/91 $25.00 © 4891 IEEE
DOI 10.1109/WCRE.2012.67

tured data over time, or extraction of semantics from software development repositories.

7) Success and failure stories of mining unstructured data, with a critical dissemination of reasons for the specific success and failure, and future steps.

III. Goals and Expected Results

The MUD workshop, through the careful selection and review of submitted workshop short papers (up to four pages), aims to provide a highly interactive platform for researchers and developers to discuss techniques for mining unstructured data and their applications. The intended outcome of this workshop is to:

1) Establish connections among the research communities that mine unstructured data, resulting in cross-fertilization of techniques and methodologies;
2) Transfer techniques and methodologies for mining unstructured data in a common framework. The produced framework will enable researchers and practitioners to select and use appropriate techniques/tools that meets their unstructured data mining needs, so that they can focus on analyzing and interpreting data.
3) Identify open problems and challenges for mining unstructured data, in order to provide the basis for a roadmap on mining unstructured data research.

IV. Format

MUD'12 is a half-day workshop, consisting of an introductory presentation, a keynote and a fishbowl panel session for semi-structured group discussions. Additionally, we invite researchers in the field to present techniques that they have adapted from NLP/IR/ML to the software engineering domain, case studies of the effective use of these techniques, and to share their experiences in mining unstructured data with participants as short papers to be presented during the workshop.

In the panel session participants will discuss the state-of-the art on mining unstructured data and will consider open research opportunities. The authors of accepted papers will present their work in slots of 15 minutes, followed by an extended 15 minute discussion.

Through the keynote, group discussions, and paper presentations, we expect to meet the MUD'12 workshop's goal of building and maintaining a community for mining unstructured data.

V. Organizers

Alberto Bacchelli obtained his Bachelor and Master's degree in Computer Science from the University of Bologna, Italy. He is currently a Ph.D. student at the University of Lugano in the Faculty of Informatics. He is working 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.

Nicolas Bettenburg received the B.Sc. and M.Sc. degree in computer science from Saarland University in 2006 and 2008. He is currently working toward the PhD degree in computer science at Queen's University (Canada) under Ahmed E. Hassan. His research interests are in mining unstructured information from software repositories with a focus on modelling the impact of developer collaboration and communication on software quality. In addition to his past work in program committees, he was the co-chair of MUD'10.

Latifa Guerrouj is a PhD Student at Ecole Polytechnique of Montreal. She received her engineering degree with honors in software engineering in 2008 and began her PhD program in 2009 under supervision of Drs. Giuliano Antoniol and Yann-Gaël Guéhéneuc. 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 Guerrouj is also interested in data mining, empirical software engineering, and search-based software engineering. In addition to serving as an external reviewer in many conferences and journals, Latifa was a student volunteer for at ICST'12 and a publicity chair for PASED (Practical Analyses of Software Engineering Data) summer school.