2010 17th Working Conference on Reverse Engineering

Workshop on Mining Unstructured Data (MUD)

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

Nicolas Bettenburg and Bram Adams Software Analysis and Intelligence Lab (SAIL)
School of Computing, Queen's University, Canada
{nicbet, bram}@cs.queensu.ca
http://sailhome.cs.queensu.ca/mud/

***Abstract*—In software development, the knowledge of developers, architects and end users is spread out across dozens of development artifacts. Historically, structured development artifacts such as source code have been the primary focus of software engineering research, but the last couple of years have seen a dramatic increase of research on unstructured data, such as free-form text requirements and specifications, mailing lists and bug reports. Mining such data is very challenging, since it typically requires dealing with natural language fragments. Research communities in information retrieval, data mining and natural language processing have explored techniques to mine unstructured data. These techniques are usually limited in scope and intended for use in specific scenarios. We feel that the knowledge gathered by these research efforts should be consolidated and propagated to the empirical software engineering communities. The MUD (Mining Unstructured Data) workshop aims to provide a highly interactive forum for researchers and developers to put challenges of, solutions for and experiences with mining unstructured data into a common reference frame and to build connections between the various communities.**

I. Motivation

Understanding software projects requires more than reverse-engineering source code. Bug reports, execution logs, mailing lists, code review reports, change logs and requirements documents each contain a significant amount of implicit developer knowledge. These documents consist of unstructured data, embedded between natural language text. Mining unstructured data is challenging, since traditional parsing and extraction techniques typically cannot handle free-form data well or identify structured components in unstructured data.

Up until now, researchers have experimented with various technologies, ranging from information retrieval techniques, such as LDA, and data mining techniques, such as hierarchical clustering, to natural language processing and ad hoc approaches. Each technique provides different ways of dealing with the complexities of unstructured data, for specific purposes. This variety makes it hard for researchers and practitioners to determine the right technique for their needs, and, once a technique has been selected, to use it effectively.

II. Topics

The MUD workshop aims to put the existing work on mining unstructured data into a common reference frame, and to identify open research challenges. The topics addressed by this workshop include, but are not limited to:

- classifying techniques for extracting unstructured data
- identifying open research challenges
- dealing with imperfect data
- evaluating the performance of MUD extractors
- cross-linking unstructured data artifacts

III. Goals and Expected Results

The MUD workshop aims to provide a highly interactive forum for researchers and developers to discuss the existing techniques for mining unstructured data, their similarities and differences. The intended outcome of this workshop is to:

1) build connections between the various communities that mine unstructured data, resulting in cross-fertilization of techniques and methodologies;
2) put existing techniques and methodologies for mining unstructured data in a common reference frame, enabling practitioners to find the right tool for their needs;
3) identify open problems and challenges for mining unstructured data, providing the basis for a roadmap on mining unstructured data research.

We plan to post the discussion output on the MUD website, enabling interested researchers to benefit from new insights. If successful, we plan to organize a second edition of the MUD workshop, with regular paper submissions, focusing on the problems and challenges identified during MUD 2010.

IV. Format

MUD 2010 is a half-day workshop consisting of an introductory presentation, a keynote and a fishbowl panel session for semi-structured group discussions. The introductory presentation situates mining unstructured data in the context of reverse-engineering, whereas the keynote speaker provides a general overview of the major accomplishments and challenges of mining unstructured data. The panel session aims to discuss the state-of-the art in

1095-1350/10 $26.00 © 2010 IEEE
DOI 10.1109/WCRE.2010.39

mining unstructured data and to find a consensus on open research opportunities. The panel will follow the so-called "fishbowl technique", in which everyone can walk up to be part of the panel, or leave the panel if one feels unable to further contribute to the discussion.

To foster more interaction in the panel session, we plan to break the ice by means of a "turbo mix-and-talk" session before the panel session. Each workshop attendee talks for two minutes to a random attendee she does not know yet, then switches partner after a signal. Balancing a keynote with group discussion is essential for meeting the MUD workshop's goal of building and maintaining a community for mining unstructured data.

V. Keynote Speaker

David Lo is an assistant professor in the School of Information Systems, Singapore Management University. His research interests include dynamic program analysis, specification mining, and pattern mining. He has worked on the extraction of behavioral models from execution logs and the analysis of textual bug reports, both of which involve mining of unstructured data. For these problems, he has investigated the use of data mining, information retrieval, and natural language processing techniques. Lo received a PhD in computer science from the National University of Singapore.

VI. Organizers

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 enriching source code with information from developer communication. In addition to his past work in program committees, he is the tool demo co-chair of WCRE 2010.

Bram Adams is an adjunct assistant professor in the SAIL lab at Queen's University (Canada). He obtained his PhD at the GH-SEL lab at Ghent University (Belgium). Bram is interested in (amongst others) reverse-engineering of build systems, aspect languages for C, crosscutting concern mining and mining mailing list data. In addition to serving in multiple program committees and being tool demo co-chair of WCRE 2009, Bram co-organized the PLATE 2009 and ACP4IS 2010 workshops.