Mining Software Engineering Data

Tao Xie
North Carolina State Univ.
USA
xie@csc.ncsu.edu

Jian Pei
Simon Fraser Univ.
Canada
jpei@cs.sfu.ca

Ahmed E. Hassan
Univ. of Victoria
Canada
ahmed@ece.uvic.ca

Abstract

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 tutorial presents the latest research in mining Software Engineering (SE) data, discusses challenges associated with mining SE data, highlights SE data mining success stories, and outlines future research directions. Participants will acquire 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.

1. Introduction

Software engineering data (such as code bases, execution traces, historical code changes, mailing lists, and bug databases) contains a wealth of information about a software project's status, progress, and evolution. Many studies have emerged that use this data to support various aspects of software development within industrial and open source settings. Working with Nokia, Gall *et al.* [4] have shown that software repositories can help developers change legacy systems by pointing out hidden code dependencies. Working with Bell Labs and Avaya, Graves *et al.* [5] and Mockus *et al.* [8] demonstrated that historical change information can support management in building reliable software systems by predicting bugs and effort. Working on open source projects, Chen *et al.* [3] have shown that historical information can assist developers in understanding large systems.

Although the idea of applying data mining techniques on software engineering data has existed since mid 1990s [7], the idea has especially attracted a large amount of interest lately within software engineering. The workshop in Mining Software Repositories (MSR) has been recognized as the most attended workshop at ICSE since 2001. MSR 2006 was oversubscribed. As a reflection of the great interest in the area and the importance of the MSR work within the context of software engineering, the best papers for three of the major conferences within SE (ICSE, ASE, and ICSM) for 2006 are on applying data mining techniques on SE data. A recent issue of IEEE Transactions on Software Engineering (TSE) on the MSR topic received over 15% of all the submissions to the TSE in 2005 [6].

The tutorial will provide participants with an overview of the field of mining software engineering data, as shown in Figure 1. In particular, the tutorial will cover the following topics along three dimensions (software engineering, data mining, and future directions):

1. **Software Engineering:**
   (a) What types of SE data are available to be mined?
   (b) Which SE tasks can be helped using data mining?
   (c) How are data mining techniques used in SE?
2. **Data Mining:**
   (a) What are the challenges in applying data mining techniques to SE data?
   (b) Which data mining techniques are most suitable for specific types of SE data?
   (c) What are freely available data mining and analysis tools (e.g., R [1] and WEKA [2])?
3. **Future Directions:** What are the challenges and opportunities for the data mining and software engineering communities?

The tutorial will cover these topics through case studies from recent software engineering conferences. Participants will gain the knowledge needed to accomplish the following tasks:

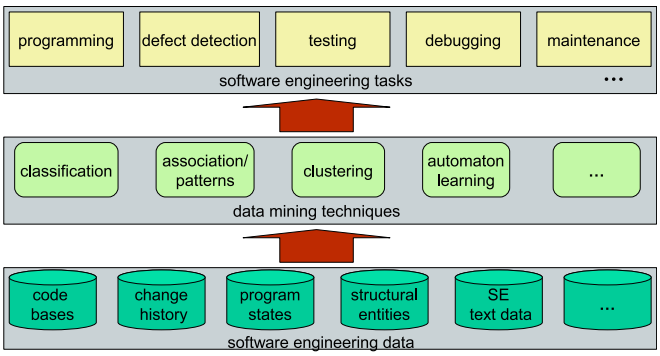

Figure 1. Overview of mining SE data

1. Appreciate the latest advancement and success stories in the field of mining SE data;
2. Conduct leading-edge research in the field of mining SE data;
3. Apply data mining techniques on their own SE data using advanced data mining analysis tools and algorithms;
4. Contrast their results relative to other work within the field;
5. Recognize open problems and possible research directions within the field.

2. Detailed Overview

The tutorial will provide a good understanding of existing research on mining SE data. The tutorial will categorize the existing research [9] in this field into three major perspectives: data sources being mined, tasks being assisted, and mining techniques being used. Figure 1 shows such a categorization with the bottom part as a set of software engineering data being mined, the middle part as a set of mining techniques being used, and the top part as a set of software engineering tasks being assisted.

From the categorization, we intend to investigate the following four issues. First, we intend to identify inherent challenges of mining software engineering data. We shall elaborate the essential requirements in software engineering, and analyze the differences between mining software engineering data and mining other types of scientific and engineering data. We shall discuss what types of data mining techniques are desired in software engineering, and how they should be customized to fit the requirements and characteristics of SE data.

Second, we intend to understand the current research and development frontier of data mining practice in software engineering. We shall summarize several kinds of data mining problems in software engineering that are under active investigation based on three major perspectives: data sources being mined, tasks being assisted, and mining techniques being used. Through this discussion, researchers can rapidly join this active research area and gain immediate access to commonly available mining techniques for real problems.

Third, we intend to analyze successful cases of mining SE data. We shall review and demonstrate briefly several research prototypes of data-mining systems for software engineering. Through the case studies, the participants can understand how to build a testbed for research and development.

Finally, we intend to give an overview on commonly used data mining tools. Our overview will help the participants gain a better understanding of available tools. The participants can use such tools in order to explore their data and integrate data mining techniques in their research and day to day work.

References

[1] The R Project for Statistical Computing. Available online at `http://www.r-project.org/`.

[2] Weka 3: Data Mining Software in Java. Available online at `http://www.cs.waikato.ac.nz/ml/weka/`.

[3] A. Chen, E. Chou, J. Wong, A. Y. Yao, Q. Zhang, S. Zhang, and A. Michail. CVSSearch: Searching through source code using CVS comments. In *Proceedings of the 17th International Conference on Software Maintenance*, pages 364–374, Florence, Italy, 2001.

[4] H. Gall, K. Hajek, and M. Jazayeri. Detection of logical coupling based on product release history. In *Proceedings of the 14th International Conference on Software Maintenance*, pages 190–198, Bethesda, Washington D.C., 1998.

[5] T. L. Graves, A. F. Karr, J. S. Marron, and H. Siy. Predicting fault incidence using software change history. *IEEE Trans. Softw. Eng.*, 26(7):653–661, 2000.

[6] A. E. Hassan, A. Mockus, R. C. Holt, and P. M. Johnson. Guest editor's introduction: Special issue on mining software repositories. *IEEE Trans. Softw. Eng.*, 31(6):426–428, 2005.

[7] M. Mendonca and N. L. Sunderhaft. Mining software engineering data: A survey. A DACS state-of-the-art report, Data & Analysis Center for Software, Rome, NY, 1999.

[8] A. Mockus, D. M. Weiss, and P. Zhang. Understanding and predicting effort in software projects. In *Proceedings of the 25th International Conference on Software Engineering*, pages 274–284, Portland, Oregon, 2003.

[9] T. Xie. Bibliography on mining software engineering data. Available online at `http://ase.csc.ncsu.edu/dmse/`.