Xie, T., Thummalapenta, S., Lo, D., & Liu, C. (2009). **Data mining for software engineering**. *Computer*, *42*(8), 55-62.

Reviewed by Roberto A. Bittencourt
Assistant Teaching Professor

University of Victoria

Summary

Main Goal

- Unleash the potential of data mining techniques used on software engineering data in order to improve software engineering tasks from the knowledge available in software system data.

Problem

- Software development produces high amounts of data that are not used to their full potential to improve typical software engineering tasks.

Solution, in short

- The authors present examples of data mining algorithms used in the context of software engineering data (sequences, graphs and text) to recover knowledge about software that may help software developers with their tasks. They also describe a general methodology for mining software engineering data with five main steps used in the process.

Detailed Solution

Data produced by software engineering with potential for mining

- Sequences
  - Execution traces
  - Static traces
  - Co-changed code locations
- Graphs
  - Dynamic call graphs
  - Static call and dependence graphs
- Text
  - Issues / Bug reports
  - Emails
  - Code comments
  - Documentation

Steps of a methodology to mine software engineering data

1. Determine SE task
2. Collect/investigate SE data
3. Preprocess data
4. Adopt/adapt/develop mining algorithm
5. Postprocess/apply mining results

Software Engineering Mining Challenges

- Requirements unique to software engineering
  - Knowledge of both software engineering and data mining
- Complex data and patterns
  - Multiple correlated and linked data types
- Large-scale data
  - Internet-scale software repositories (open source world)
- Just-in-time mining
  - Providing answers to software developers on their ongoing software project

Mining Examples

- Sequences
  - Iterative pattern mining (finding behavioral code patterns)
  - Temporal rule mining (finding temporal invariants)
  - Sequence diagram and FSM mining (recovering documentation)
  - Sequence association rule mining (finding exception handling rules)
- Graphs
  - Discriminative graph mining (finding potential bugs from testing)
  - Graph classification (finding potential bugs from testing)
- Text
  - Finding bug report duplicates using NLP techniques
  - Combining bug report info with execution traces to find duplicates

Challenges to future research

- Adapting general data mining techniques to the context of software engineering
- Expanding the scope of software engineering tasks that can benefit from data mining
- Need for increased scalability of mining algorithms for use in SE tools to perform SE tasks

Critical Evaluation

Positioning on the text

- In general, I agree with the authors on the potential of mining software engineering data to improve the work (both in terms of quality and productivity) of software developers.
- I think the authors are sometimes unclear on explaining the mining algorithms and the software engineering tasks themselves, which makes it hard to buy the data mining package from this paper alone

Pros

- Description of different software data types to be mined
- Description of a general approach to mine software engineering data
- Description of the challenges to mine software engineering data
- Examples of data mining algorithms used in the context of software engineering tasks
- Examples of data mining come from real open-source software systems, which leads to potential application of the techniques on those systems (e.g., documentation)

Cons

- Too many "superficial" examples may not allow a deeper understanding of data mining usage to improve software development work
- Some examples illustrate the algorithms but are hard to catch since the software engineering task is not usually the starting point
- The examples of graph mining, figures included, do not start from a software engineering task, and hard to understand

So what?

- I think the authors were fortunate to describe some important issues in data mining for software engineering: data, methodology, challenges, and data mining algorithms
- I think the authors missed an opportunity to depart from software engineering tasks
- The lack of further discussion of recommender systems (they do a little though) for software engineering leaves an important gap on the potential use of data mining for software engineering
- Data mining for software engineering has become one of the main empirical research techniques in software engineering research, and this paper marks a turning point

Xie, T., Thummalapenta, S., Lo, D., & Liu, C. (2009). **Data mining for software engineering**. *Computer*, *42*(8), 55-62.

Reviewed by Roberto A. Bittencourt
Assistant Teaching Professor

University of Victoria