Mining Software Repositories

Session 1

Infrastructure and extraction

Discussion Leader: Daniel M. German

The Stages

1. Data Extraction
2. Data Mining/Facts Finding/Change Patterns/System Understanding
3. Integration and Presentation

The Extraction Stage

- The *dirty* work, but somebody has to do it
- Lots of *raw* data out there
  - Usually Open Source
  - Difficult to gain access to Closed source data

The Issues

- Why do we need extract historical data?
- Without a purpose, this data might have no value

The Issues...

- What to extract? (*software trails*)
  - Code
    - Releases
    - Versioning history
  - Defects
  - Documentation
    - Explicit (man pages, help system, design documents)
    - Implicit (email messages)
    - Web site

The Issues...

- From Where
  - What projects to select?
  - The software process might have an impact in the way the historical data gets recorded
  - It is necessary to understand this process
  - Different projects store data in different ways

The Papers

- The Perils and Pitfalls of Mining SourceForge
  by *James Howison and Kevin Crowston*
- Their experiences mining sourceForge
- What they learnt spidering the site
- Some potential mistakes in the analysis of the extracted data

The Papers...

- Text is Software Too by *Alexander Dekhtyar, Jane Huffman Hayes and Tim Menzies*
- Mining of textual requirements documents
- “Text mining from software engineering text is a hight risk, high return adventure.”

The Papers...

- Mining CVS Repositories, the softChange experience by *Daniel German*

- The revision history of the source code says a lot about the project:
  - it highlights the process, the architecture evolution, hidden relationships between files...

- The Concurrent Versions System (CVS) is a major source of historical data

The Papers

- Research Infrastructure for Empirical Science of F/OSS
  by *Les Gasser, Gabriel Ripoche and Robert Sandusky*

- Preprocessing CVS Data for Fine-Grained Analysis
  by *Thomas Zimmerman and Peter Weissgerber*

Discussion: the Issues, revisited

- Several people are working in the same problems
  - Comparison?
  - Collaboration? (Avoid reinventing the wheel)
- Nomenclature?
- Choosing projects for analysis?
- Sharing data?
- Sharing the extractors?