Session 2:
Integration and Presentation

Katsuro Inoue
Osaka University

Papers in this Session

1. GlueTheos: Automating the Retrieval and Analysis of Data from Publicly Available Software Repositories
2. Using CVS Historical Information to Understand How Students Develop Software
3. Database Techniques for the Analysis and Exploration of Software Repositories
4. Empirical Project Monitor: A Tool for Mining Multiple Project Data

1. GlueTheos: Automating the Retrieval and Analysis of Data from Publicly Available Software Repositories

- Objective:
  - Analysis of free software systems
  - Measurement of LOC and its visualization
- Approach:
  - Script-based analyzer for CVS
- Strength:
  - Flexible architecture
- Weakness:
  - Applicability to other data ?

2. Using CVS Historical Information to Understand How Students Develop Software

- Objective:
  - Analysis of activities in a development team
- Approach:
  - Hierarchical analysis (file-individual-team) of CVS data
- Strength:
  - Fine granularity analysis -> activity observation
- Weakness:
  - Scalability to larger projects ?

3. Database Techniques for the Analysis and Exploration of Software Repositories

- Objective:
  - Analysis of e-mail archive
- Approach:
  - Putting everything into a single SQL DB and mining the DB
- Strength:
  - Once DB has been created, every operation is performed on it
- Weakness:
  - Performance ?
  - Data format translation to SQL DB

4. Empirical Project Monitor: A Tool for Mining Multiple Project Data

- Objective:
  - Real-time process monitor
- Approach:
  - CVS, Mailman, and GNATS data to standardized XML database
- Strength:
  - data format standardization
- Weakness:
  - Applicability to other data sources

General Model behind the Works

Feedback

Projects

- Software Companies
- Open Source Projects

Collection

Session 1

Integration

Analysis

Presentation

Session 3, 4, 5

Integration

1. GlueTheos: Raw CVS data
   (+ XML, SQL for external interface)
2. Student Activities: CVS logs
3. Database: SQL DB
4. EPM: XML Standardized data

Presentation

1. GlueTheos: Graphical presentation of project's LOC
2. Student Activities: Graphical presentation of detail activities
3. Database: Web-based browser for mail clusters
4. EPM: Graphical presentation of LOC, # of mails and bugs

Issues on Integration

- Standardize data format
  - Useful for data exchange?
  - Translation overhead
- Database v.s. Raw files (CVS, Mail, ...)
  - Easiness of data mining
  - Translation overhead and performance

Issues on Presentation

- Types of presentation
  - Based on the goal of measurement/analysis
  - Currently, most works provide simple metrics graphs
  - Other analyses -> different presentations

    ex. 3. mail analysis -> cluster browser
- Other types of repository analysis ?
  - Reuse, knowledge share, ...