Software Comprehension and Reuse supported by Analytics and Repository Mining

Marcelo de Almeida Maia
Computer Science Department
Federal University of Uberlândia
marcmaia@facom.ufu.br

SEIF Brazil Workshop 2013
A little bit about Uberlândia, UFU, LASCAM

- Uberlândia: Strategic location
  - Population: 700,000
  - Reference in Logistics
- UFU
  - Relatively young institution
- Graduate Program in Computer Science
  - 2001 MSc  2011 PhD
- This talk is around giving shape to
  - Laboratory of Software Comprehension, Analytics and Mining
    - 3 PhD students
    - 4 MSc students
    - 2 undergraduate students
- Agenda
  - Motivation
  - Some previous/current work
  - Some related work
  - Our vision
How much does open source code worth?

In 2009, in a Computerworld column:

LOC (billions)

- In 2009, OSS estimated in ~US$ 400 bi

Source: Computerworld / www.ohloh.net
OSS and Software Development

Few examples

- **Infrastructure**
  - DBMS: Postgres
  - Application Servers: Jboss (Wildfly)
  - Cloud Platforms: CloudStack
  - Distributed Computing: Hadoop

- **Libraries and Frameworks**
  - Search engines: Lucene
  - Frameworks: Spring

- **IDEs and SE Tools**
  - Eclipse, Netbeans
  - SonarQube
OSS and Software Development

Few examples

- Infrastructure
  - DBMS: Postgres
  - Application Servers: JBoss, (WebLogic)
  - Cloud Platforms: OpenStack
  - Distributed Computing: Hadoop
- Libraries and Frameworks
  - Search Engines: Lucene
  - Frameworks: Spring
- IDEs and SE Tools
  - Eclipse, Netbeans
  - SonarQube

GREAT!

But, this is not the treasure we’re seeking!

At least not as a box!

*If ever exists! 😊
What is the “treasure”? 

- “Raw” LOC !??
  White-box Reuse !??
  Fine-grained code elements !??

- Treasure or Trash? Eye of the beholder...
What is the “treasure”? 

- Raw LOC!!!
  - White-box Reuse !!!
  - Fine-grained code elem

- Treasure or Trash?

60% of structure
90% of furniture = Trash
Is it worth?

### Gain
- Internal algorithms
- Internal concepts
- Templates
- Partial structure
- Architecture

### Pain
- Coupling impact
- Lack of high-level doc
- Context understanding
- Quality assessment
- Adaptation/Restructuring...
What is the “treasure”?

- Code or People?

How can we reuse the crowd knowledge?
Some of our past work

Execution traces

- Using execution traces for comprehension and maintenance
  - Difficulties for large scale analysis of repositories

Some of our past work

Characterization of domains, PL constructions evaluation

- Difficulties for cause-effect understanding, measurement extraction for features.

Coupling x Categories (MSR2013)    FOP x SPLs (SCP2014)
Some of our current work

Version history and Traces

- **Modularity Assessment and Comprehension**
  - Difficulties for cause-effect understanding, ground-truth evaluation
  - Necessity to have large history of changes and high coverage traces

Co-change clusters (submitted)
Some of our current work

Q&A mining

- Structured Information Extraction for Recommendation Systems using machine learning and IR (WIP)
- Subjective nature of unstructured data

Post classification based on intent

Cookbook extraction

```
19 : user list menu page item web select display gray convert a JTF-Geometry into an AWT-Shape?

17 : paint draw method graphic common call object s a Java Component in an image file event not called

16 : mouse click event press want listen user mouse automatically move from one TextField to other

15 : how to fire tab key event

The following example shows how to simulate mouse and key presses in Java:

```
Some of related state-of-the-art work
Poshyvanyk and colleagues

- ASE Conf. 2013

**Detecting Bad Smells in Source Code Using Change History Information**

- Necessity to have large history of changes

**Fig. 1.** HIST: The proposed code smell detection process.
Some of related state-of-the-art work
Lanza and colleagues

- CSMR2013 - Seahawk
  - formulates queries from the current context in the IDE, and presents a ranked and interactive list of results
  - Low NDCG: 9.07%

Figure 1: Seahawk User Interface
Some of related state-of-the-art work
Kakarontzas and colleagues

- JSS2013 – FWBR Metric (Facilitating White-Box Reuse)
  - Based on D-layer... regression D-layer -> CK function
  - Lack of external validation
Envisioning our goals
Gluing the pieces together

<table>
<thead>
<tr>
<th>Code Changes</th>
<th>Issues</th>
<th>Traces</th>
<th>Posts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Warehouse: Extraction, Transform, Load</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Data Analysis: Statistics, Data Mining, Machine Learning, IR</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Recommendation Systems Integrated into IDEs</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Reuse in Repos</strong></td>
<td><strong>Modular Rejuvenation</strong></td>
<td><strong>Contextual How-tos</strong></td>
<td></td>
</tr>
</tbody>
</table>
At last, but not least
Acknowledgement to MS Research for SEIF-Brazil

<table>
<thead>
<tr>
<th>Code</th>
<th>Changes</th>
<th>Issues</th>
<th>Traces</th>
<th>Posts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Warehouse: Extraction, Transform, Load</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Analysis: Statistics, Data Mining, Machine Learning, IR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Visual Studio
  - Reuse from Codeplex
  - Modular Rejuvenation
  - Contextual How-tos
QUESTIONS????

Software Comprehension and Reuse supported by Analytics and Repository Mining

Marcelo de Almeida Maia

Computer Science Department
Federal University of Uberlândia

marcmaia@facom.ufu.br