The Road Ahead for Mining Software Repositories

Ahmed E. Hassan
Queen’s University
Canada
Mining Software Repositories (MSR)

- Transforms static record-keeping repositories to **active** repositories
- Makes repos data **actionable** by uncovering hidden **patterns** and **trends**
MSR researchers analyze and cross-link repositories

New Bug Report
- Estimate fix effort
- Mark duplicates
- Suggest experts and fix
MSR researchers analyze and cross-link repositories

New Change

Suggest APIs
Warn about risky code or bugs
Suggest locations to co-change

3-6x
Performance over only static dependencies
(Recall 78%, Precision 64%)
Supporting software understanding (NETBSD)

Conceptual (proposed)

- Hardware Trans.
- Kernel Fault Handler
- Pager
- Virtual Addr. Maint.
- VM Policy
- FileSystem

Concrete (reality)

- Hardware Trans.
- Kernel Fault Handler
- Pager
- Virtual Addr. Maint.
- VM Policy
- FileSystem

Mining supports software understanding (NETBSD)

- Eight unexpected dependencies
- All except two dependencies existed since day one:
  - Virtual Address Maintenance $\rightarrow$ Pager
  - Pager $\rightarrow$ Hardware Translations

**Table:**

<table>
<thead>
<tr>
<th>Which?</th>
<th>vm_map_entry_create (in src/sys/vm/Attic/vm_map.c) depends on pager_map (in /src/sys/uvm/uvm_pager.c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who?</td>
<td>cgd</td>
</tr>
<tr>
<td>When?</td>
<td>1993/04/09 15:54:59 Revision 1.2 of src/sys/vm/Attic/vm_map.c</td>
</tr>
<tr>
<td>Why?</td>
<td>from sean eric fagan: it seems to keep the vm system from deadlocking the system when it runs out of swap + physical memory. prevents the system from giving the last page(s) to anything but the referenced &quot;processes&quot; (especially important is the pager process, which should never have to wait for a free page).</td>
</tr>
</tbody>
</table>
Opportunities in the Road Ahead

• Going beyond code and bugs
• Taming the complexity of MSR
• Showing the value of repositories
• Easing the adoption of MSR
Opportunities in the Road Ahead

Going beyond code and bugs

- **Explore non-structured data**
  - Social aspects: emails and comments
- **Link data between repos**
- **Seek non-traditional repos**
  - Demonstrate the value of IDE interactions or build failures repos
- **Understand the limitation of repos**
  - Causation vs. Correlation
    - Small number of committers in OS projects

MSR 2004-2008: ~80% of publications focus on code and bugs
Opportunities in the Road Ahead

Repository

Extract

Analyze

Adopt

Results

• Simplify the extraction of high quality data

Taming the complexity of MSR

main() {
    int a;
    /*call
     *help*/
    helpInfo();
}

helpInfo() {
    errorString!
}

main() {
    int a;/*call
     *help*/
    helpInfo();
}

V1: Undefined func. (Link Error)

V2: Syntax error

V3: Valid code

- Toolkits and extracted data (e.g. FLOSSMetrics) are needed
- Heuristics should be empirically verified
- Acknowledgement mechanism needed for extractors

• Deal with skew in repository data
  - Visualization can help spot skew
  - Guidelines and re-sampling/robust techniques are needed

• Improve the quality of repository data
  - Provide tools for annotation of repos data at creation
Opportunities in the Road Ahead

- **Simplify the extraction of high quality data**
  - Toolkits and extracted data (e.g. FLOSSMetrics) are needed
  - Heuristics should be empirically verified
  - Acknowledgement mechanism needed for extractors

- **Deal with skew in repository data**
  - Visualization can help spot skew
  - Guidelines and re-sampling/robust techniques are needed

- **Improve the quality of repository data**
  - Provide tools for annotation of repos data at creation
Opportunities in the Road Ahead

- **Understand the needs of practitioners**
  - Predicting buggy modules:
    - Buggy modules are well-known 😞
    - Predicting fault occurrences at module level is too coarse
- **Study the performance in practice**
  - Tools affecting the repos data
- **Show the practical benefits**
  - Statistical improvements not sufficient
  - Cost of maintenance should be evaluated
- **Evaluate on non-open source systems**
Opportunities in the Road Ahead

- **Simplify access to techniques**
  - Integration into IDEs (HATARI, Hipikat, Myln, eRose)
  - A web service demonstration for an open source project
    - A continuously updating MSR Challenge

- **Help practitioners make decisions**
  - MSR should aim to support not replace practitioners
Mining Software Repositories

Mining Software Repositories (MSR)
- Transforms static record-keeping repositories to active repositories
- Makes repos data actionable by uncovering hidden patterns and trends

Supporting Software Understanding (NETBSD)

Opportunities in the Road Ahead
- Going beyond code and bugs
- Taming the complexity of MSR
- Showing the value of repositories
- Easing the adoption of MSR