Risky Files: An Approach to Focus Quality Improvement Effort

Audris Mockus    Randy Hackbarth    John Palframan

Avaya Labs Research
211 Mt Airy Rd
Basking Ridge, NJ 07920
audris@avaya.com

Aug 21, 2013
Motivation

Make quality of the code transparent

Indications

- Development transferred
- Few original authors remain
- A long development history
- Many customers/customer issues
- A component of many projects
Motivation

Make *quality* of the code *transparent*

Indications

- Development transferred
- Few original authors remain
- A long development history
- Many customers/customer issues
- A component of many projects
Motivation

Make *quality* of the code *transparent*

Indications

- Development transferred
- Few original authors remain
- A long development history
- Many customers/customer issues
- A component of many projects
Benefits

Top 1% of all files contribute to 60+% of field defects

Make Transparent

- Where to rebuild lost expertise
- Where to focus quality improvement

Provide guidance for

- Cost effective actions
- Practices to reduce future defects
Approach Outline

- Data processing
  - Accessing data sources
  - Linking data sources
  - Obtaining risk predictors
- Prioritized candidate list
  - Details needed for action
    - Related files
    - Modification Requests (MRs)
    - Customer Reported Defects (CFDs)
    - Developer expertise
  - Determine and schedule actions
- Monitor actions and measure quality improvement
Data Sources

- Code changes
  - 1K+ projects using git/svn/ClearCase/SCCS/other VCS
  - 50M+ changes
Data Sources

- Code changes
  - 1K+ projects using git/svn/ClearCase/SCCS/other VCS
  - 50M+ changes

- MRs: Why change was made?
  - ClearQuest/JIRA/other: 1.6M MRs
Data Sources

- Code changes
  - 1K+ projects using git/svn/ClearCase/SCCS/other VCS
  - 50M+ changes

- MRs: Why change was made?
  - ClearQuest/JIRA/other: 1.6M MRs

- Support: which MRs came from users (CFDs)?
  - Customer support (Siebel)

- Directory: who represents that login?
  - Corporate directory
  - Yellow pages to map login to corporate handle
Linking Data

- MRs from code commit comments
- Corporate handle for commit login
- CFDs from Siebel
Linking Data

- MRs from code commit comments
- Corporate handle for commit login
- CFDs from Siebel

- Identify related (copied in the past) files
  - $f_1$ is directly related ($\sim$) to $f_2$ if
    $\exists v_1, v_2 : f_1(v_1) = f_2(v_2)$
    where $f(v)$ is a string representing version $v$ of file $f$
  - $f_1$ is related to $f_2$ (a transitive closure of $\sim$) iff
    $\exists F_1, \ldots, F_k : f_1 \sim F_1, F_1 \sim F_2, \ldots, F_k \sim f_2$
io.c ~ fio.c: directly related files
Determine risk factors most strongly associated with future customer-reported defects

Identify and prioritize files (equivalence classes)

- Risk predictors
  - Number of changes, CFDs
  - Number of authors, number who left
  - Size in LOC
  - Author experience
  - Number of static analysis warnings
  - % test coverage

- Risk prioritization
  - Fit a logistic regression model
  - Use a minimal subset to prioritize

- Produce top 1% risky file report
Determine risk factors most strongly associated with future customer-reported defects

Identify and prioritize files (equivalence classes)

- Risk predictors
  - Number of changes, CFDs
  - Number of authors, number who left
  - Size in LOC
  - Author experience
  - Number of static analysis warnings
  - % test coverage

- Risk prioritization
  - Fit a logistic regression model
  - Use a minimal subset to prioritize

- Produce top 1% risky file report
For subject matter experts (SMEs)

- In three formats
  - Hypertext, sortable by metrics, CSV
- Hypertext: for each file
  - Link to related files
  - Two most recent CFDs
  - Link to MRs
  - Link to authors/experience
  - Relevant metrics: LOC, coverage, ...
- Checklist of suggested actions
Example: Risky File Author View

candidate risky file list

Format 1 - Example of Login Page

Authors of `dca/one/branche/merge-ui/application/WindowLaunchPad.xaml.cs`

<table>
<thead>
<tr>
<th>Login</th>
<th>Name</th>
<th>email</th>
<th>Phone</th>
<th>NDelta (this file)</th>
<th>TotDelta (all files)</th>
<th>From (all files)</th>
<th>To (all files)</th>
</tr>
</thead>
<tbody>
<tr>
<td>admin</td>
<td>Christopher</td>
<td><a href="mailto:admin@na-west.exchange.avaya.com">admin@na-west.exchange.avaya.com</a></td>
<td>+1 908 880 8888</td>
<td>642</td>
<td>34544</td>
<td>2007-07-26</td>
<td>2013-04-12</td>
</tr>
<tr>
<td>alias</td>
<td>Adam</td>
<td><a href="mailto:ali@apac.exchange.avaya.com">ali@apac.exchange.avaya.com</a></td>
<td>+1 408 999 9999</td>
<td>351</td>
<td>10820</td>
<td>2008-07-22</td>
<td>2011-10-21</td>
</tr>
<tr>
<td>kathleen</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>sasha1</td>
<td>Sasha</td>
<td><a href="mailto:sasha@mera.ru">sasha@mera.ru</a></td>
<td>+7 8310990988</td>
<td>121</td>
<td>16799</td>
<td>2012-02-03</td>
<td>2013-04-01</td>
</tr>
<tr>
<td>sasha2</td>
<td>Sasha</td>
<td><a href="mailto:sasha@mera.ru">sasha@mera.ru</a></td>
<td>+7 8310990988</td>
<td>115</td>
<td>4176</td>
<td>2010-09-05</td>
<td>2013-04-16</td>
</tr>
<tr>
<td>natalia</td>
<td>Natalia</td>
<td><a href="mailto:natalia@mera.ru">natalia@mera.ru</a></td>
<td>+7 8310990988</td>
<td>88</td>
<td>4858</td>
<td>2010-08-09</td>
<td>2013-01-15</td>
</tr>
<tr>
<td>jennifer</td>
<td>Jennifer</td>
<td><a href="mailto:jennifer@na-west.exchange.avaya.com">jennifer@na-west.exchange.avaya.com</a></td>
<td>+1 303 880 8888</td>
<td>70</td>
<td>910</td>
<td>2011-11-10</td>
<td>2013-03-27</td>
</tr>
<tr>
<td>anna</td>
<td>Anna</td>
<td><a href="mailto:anna@na-west.exchange.avaya.com">anna@na-west.exchange.avaya.com</a></td>
<td>+1 303 880 8888</td>
<td>68</td>
<td>236143</td>
<td>2010-09-03</td>
<td>2012-07-19</td>
</tr>
</tbody>
</table>

Link to Post Entry of Login:
No link == No longer with Avaya

Information about the user across all file in all the repositories scanned
Expert assignment and training

- Use file authorship to determine/assign SME
- SME is trained how to use the report and checklist
- SME examines the report to:
  - Determine action for each risky file
  - Schedule the action
SME Recommendations

- No action required if
  - E.g., will become unused; just changed with a risky file
- Control if
  - E.g., little active development in the future
- Control examples
  - Extra review SME+Owner, and testing for any change
  - If many authors: create a brief design/test guide
- Restructure if
  - Development in the future and the file is too fragile
- If no remaining authors: assign a file owner
Update on status

- Created candidate sets of risky files for 45 projects.
- Held training sessions with 17 of these projects.
- 7 of these projects are defining actions.
Discussion

- Use of Big Data
  - To make quality visible to multiple stakeholders

- Enable SMEs to take action
  - By (usually) justifying their intuition
  - By providing quantitative evidence for management
Discussion

- A patchwork on cutting-edge techniques
  - Data mining
  - Risk prediction
  - Expertise browser (link code and people)
  - Relationship among files in different repositories

- Feedback from early users
  - Need to show or drill-down to detail: code, MRs, people
  - Multiple forms of presentation
  - Role-specific aggregation
  - Bug in another project: DILLIC/DILLIGAD?