Using Data Fusion and Web Mining to Support Feature Location in Software

Meghan Revelle, Bogdan Dit, Denys Poshyvanyk

18th IEEE International Conference on Program Comprehension (ICPC’10)
Feature: a requirement that user can invoke and that has an observable behavior.
Textual Feature Location

- **Information Retrieval (IR)**
  - Searching for documents or within docs for relevant information

- **First used for feature location by Marcus et al. in 2004**.
  - Latent Semantic Indexing** (LSI)

- **Utilized by many existing approaches**: PROMESIR, SITIR, HIPIKAT etc.


Applying LSI to Source Code

- **Corpus creation**
  - Choose granularity

- **Preprocessing**
  - Stop word removal, splitting, stemming

- **Indexing**
  - Term-by-document matrix
  - Singular Value Decomposition

- **Querying**
  - User-formulated

- **Generate results**
  - Ranked list
Dynamic Feature Location

Software Reconnaissance*

Feature Invoked

Feature Not Invoked

Scenario-based Probabilistic Ranking (SPR)**


## Hybrid Feature Location

<table>
<thead>
<tr>
<th></th>
<th>PROMESIR*</th>
<th>SITIR**</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSI score</td>
<td>SPR score</td>
<td>Execution Trace</td>
</tr>
<tr>
<td>m₁₅ 0.91</td>
<td>m₅₂ 0.80</td>
<td>m₆ 0.715</td>
</tr>
<tr>
<td>m₁₆ 0.88</td>
<td>m₄₇ 0.66</td>
<td>m₄₇ 0.70</td>
</tr>
<tr>
<td>m₂ 0.85</td>
<td>m₆ 0.64</td>
<td>m₅₂ 0.70</td>
</tr>
<tr>
<td>m₆ 0.79</td>
<td>m₂ 0.53</td>
<td>m₂ 0.69</td>
</tr>
<tr>
<td>m₄₇ 0.74</td>
<td>m₁₅ 0.37</td>
<td>m₁₅ 0.64</td>
</tr>
<tr>
<td>m₅₂ 0.60</td>
<td>m₁₆ 0.34</td>
<td>m₁₆ 0.61</td>
</tr>
<tr>
<td>... ...</td>
<td>... ...</td>
<td>main</td>
</tr>
</tbody>
</table>

*Probabilistic Ranking of Methods Based on Execution Scenarios and Information Retrieval*


**Single Trace and Information Retrieval**

Data Fusion Example

Global Positioning System (GPS)
- Discrete measurements
- Meter accuracy
- Noisy
+ No drift

Inertial Navigation System (INS)
+ Continuous measurements
+ Centimeter accuracy
+ Low noise
- Drifts over time

Actual Position
Data Fusion for Feature Location

• Combining information from multiple sources will yield better results than if the data is used separately
  - Previous
    • Textual, Dynamic, and Static
  - Current
    • Textual info from IR
    • Execution info from dynamic tracing
    • Web mining
Web Mining

![Diagram](image)

**Results 1 - 10 of about 19,800,000 for web mining** (0.37 seconds)

- the free encyclopedia
  - a collection of data mining techniques to discover patterns from the targets, **web mining** can be divided into...
- Caching - Web structure mining
- Cached - Similar -

**Ins and Techniques - Google Books Results**
- Computers - 427 pages
- and related applications in a manner that encourages additional abstraction of information overflow, which...
- bn=1591404142...

**Search: A Survey**
- Google - Quick View
- Related articles
- a converging research area from several research communities, mining and when comparing research in this...
- c/download;jsessionid...?doi=10.1.1...

Upon in data mining terms, can be said to have three operations grouping of users, ...
- Cached - Similar -

**reb-mine/**
Web Mining Algorithms

PageRank

- Measure the relative importance of a web page
- Used by the Google search engine
- Link from X to Y means a vote by X for Y
- A node’s PageRank depends on # incoming links and the PageRank of nodes that link to it

Image source: http://en.wikipedia.org/wiki/Pagerank

Web Mining Algorithms

HITS

- Hyperlinked-Induced Topic Search
- Identifies hub and authority pages
- Hubs point to many good authorities
- Authorities are pointed to by many hubs

Probabilistic Program Dependence Graph*

PPDG
- Derived from feature-specific trace
- Binary weights
- Execution frequency weights

Incorporating Web Mining with Feature Location
## Feature Location Techniques Evaluated

<table>
<thead>
<tr>
<th>LSI &amp; Dynamic Analysis</th>
<th>Web Mining</th>
<th>LSI, Dyn, &amp; PageRank</th>
<th>LSI, Dyn, &amp; HITS</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSI</td>
<td>PR(bin)</td>
<td>LSI+Dyn+PR(bin)(^\text{top})</td>
<td>LSI+Dyn+HITS(h,bin)(^\text{top})</td>
</tr>
<tr>
<td>LSI+Dyn</td>
<td>PR(freq)</td>
<td>LSI+Dyn+PR(bin)(^\text{bottom})</td>
<td>LSI+Dyn+HITS(h,req)(^\text{top})</td>
</tr>
<tr>
<td>(baseline)</td>
<td>HITS(h, bin)</td>
<td>LSI+Dyn+PR(freq)(^\text{top})</td>
<td>LSI+Dyn+HITS(a,bin)(^\text{top})</td>
</tr>
<tr>
<td></td>
<td>HITS(h, freq)</td>
<td>LSI+Dyn+PR(freq)(^\text{bottom})</td>
<td>LSI+Dyn+HITS(a,req)(^\text{top})</td>
</tr>
<tr>
<td></td>
<td>HITS(a, bin)</td>
<td></td>
<td>LSI+Dyn+HITS(a,bin)(^\text{bottom})</td>
</tr>
<tr>
<td></td>
<td>HITS(a, freq)</td>
<td></td>
<td>LSI+Dyn+HITS(a,req)(^\text{bottom})</td>
</tr>
</tbody>
</table>

**Use LSI to rank methods, prune unexecuted**

**Use web mining algorithm to rank methods.**

**Use LSI to rank methods. Prune unexecuted. Use web mining algorithm to also rank methods and prune top- or bottom- ranked methods from LSI+Dyn’s results.**
Feature Location Techniques Explained

LSI+Dyn

PR(bin)

HITS(h, bin)

Source Code

LSI Query Tracer Scenario

Ranked Methods Executed Methods

Web Mining

Ranked Methods+

Ranked, Executed Methods+

Final Results
Subject Systems

• Eclipse 3.0
  - 10K classes, 120K methods, and 1.6 million LOC
  - 45 features
  - **Gold set**: methods modified to fix bug
  - **Queries**: short description from bug report
  - **Traces**: steps to reproduce bug
Bug 66914 - [typing] Error Message after undo copy/paste

Status: VERIFIED FIXED
Product: JDT
Component: Text
Version: 3.0
Platform: PC All

Importance: P2 major (vote)
Target Milestone: 3.0 RC3
Assigned To: Tom Hofmann
QA Contact:

See Also:

Attachments

<table>
<thead>
<tr>
<th>error log (5.43 KB, text/plain)</th>
<th>no flags</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004-06-14 08:17 EDT, Ralf Schmauder</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LinkedModeUI.diff (4.59 KB, patch)</th>
<th>no flags</th>
<th>Details</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004-06-18 09:59 EDT, Tom Hofmann</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Add an attachment (proposed patch, testcase, etc.) View All

Ralf Schmauder 2004-06-14 08:16:24 EDT

- create a new Class and generate the main method
- type "sysout" and use the code completion
- type the double quote
- paste Hello World into the double quotes
- try to undo without saving using Ctrl+z

using undo in the menubar does work
Subject Systems

- **Rhino 1.5**
  - 138 classes, 1,870 methods, and 32,134 LOC
  - 241 features
  - **Gold set**: Eaddy et al.’s dataset*
  - **Queries**: description in specification
  - **Traces**: test cases

## Size of Traces

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Max</th>
<th>25%</th>
<th>Med</th>
<th>75%</th>
<th>σ</th>
<th>μ</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Eclipse</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methods</td>
<td>88K</td>
<td>1.5MM</td>
<td>312K</td>
<td>525K</td>
<td>1MM</td>
<td>666K</td>
<td>406K</td>
</tr>
<tr>
<td>Unique Methods</td>
<td>1.9K</td>
<td>9.3K</td>
<td>3.9K</td>
<td>5K</td>
<td>6.3K</td>
<td>5.1K</td>
<td>2K</td>
</tr>
<tr>
<td>Size-MB</td>
<td>9.5</td>
<td>290</td>
<td>55</td>
<td>98</td>
<td>202</td>
<td>124</td>
<td>83</td>
</tr>
<tr>
<td>Threads</td>
<td>1</td>
<td>26</td>
<td>7</td>
<td>10</td>
<td>12</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td><strong>Rhino</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methods</td>
<td>160K</td>
<td>12MM</td>
<td>612K</td>
<td>909K</td>
<td>1.8MM</td>
<td>1.8MM</td>
<td>2.3MM</td>
</tr>
<tr>
<td>Unique Methods</td>
<td>777</td>
<td>1.1K</td>
<td>870</td>
<td>917</td>
<td>943</td>
<td>912</td>
<td>54</td>
</tr>
<tr>
<td>Size-MB</td>
<td>18</td>
<td>1,668</td>
<td>71</td>
<td>104</td>
<td>214</td>
<td>210</td>
<td>273</td>
</tr>
<tr>
<td>Threads</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Research Questions

• RQ1
  - Does combining web mining algorithms with an existing approach to feature location improve its effectiveness?

• RQ2
  - Which web-mining algorithms, HITS or PageRank, produces better results?
Data Collection & Testing

- **Effectiveness measure**
  - Descriptive statistics
    - 45 Eclipse features
    - 241 Rhino features

- **Statistical Testing**
  - Wilcoxon rank sum test
  - Null hypothesis
    - There is no significant difference between the effectiveness of \( X \) and the baseline (LSI+Dyn).
  - Alternative hypothesis
    - The effectiveness of \( X \) is significantly better than the baseline (LSI+Dyn).

<table>
<thead>
<tr>
<th>LSI score</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( m_{15} )</td>
<td>0.91</td>
</tr>
<tr>
<td>( m_{16} )</td>
<td>0.88</td>
</tr>
<tr>
<td>( m_2 )</td>
<td>0.85</td>
</tr>
<tr>
<td>( m_6 )</td>
<td>0.79</td>
</tr>
<tr>
<td>( m_{47} )</td>
<td>0.74</td>
</tr>
<tr>
<td>( m_{52} )</td>
<td>0.60</td>
</tr>
</tbody>
</table>

Effectiveness = 4
Results: Web Mining Techniques

Eclipse

Rhino

- LSI
- LSI+Dyn
- PR(freq)
- PR(bin)
- HITS(a, freq)
- HITS(a, bin)
- HITS(h, freq)
- HITS(h, bin)
Results: IR, Dyn, & Web Mining

Eclipse

1. LSI+Dyn
2. LSI+Dyn+PR(freq)\text{top} \{40, 60\}%
3. LSI+Dyn+PR(freq)\text{bot} \{20, 70\}%
4. LSI+Dyn+PR(bin)\text{top} \{40, 60\}%
5. LSI+Dyn+PR(bin)\text{bot} \{10, 70\}%
6. LSI+Dyn+HITS(a, freq)\text{top} \{30, 70\}%

Rhino

7. LSI+Dyn+HITS(a, freq)\text{bot} \{40, 60\}%
8. LSI+Dyn+HITS(h, freq)\text{top} \{10, 70\}%
9. LSI+Dyn+HITS(h, freq)\text{bot} \{60, 50\}%
10. LSI+Dyn+HITS(a, bin)\text{top} \{20, 70\}%
11. LSI+Dyn+HITS(a, bin)\text{bot} \{40, 40\}%
12. LSI+Dyn+HITS(h, bin)\text{top} \{10, 70\}%
13. LSI+Dyn+HITS(h, bin)\text{bot} \{70, 60\}%
A Case in Point: Eclipse exclusion filter

LSI = 1,696

LSI+Dyn+ 
HITS(h, bin)\text{bottom} 
= 24
Results of the Wilcoxon Rank Sum test comparing these techniques to the baseline, LSI+Dyn.

Null Hypothesis: There is no significant difference between the effectiveness of X and the baseline, LSI+Dyn.

<table>
<thead>
<tr>
<th></th>
<th>Eclipse</th>
<th>Rhino</th>
<th>Null Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>PR(bin)</td>
<td>1</td>
<td>1</td>
<td>Not Rejected</td>
</tr>
<tr>
<td>PR(freq)</td>
<td>1</td>
<td>1</td>
<td>Not Rejected</td>
</tr>
<tr>
<td>HITS(h, bin)</td>
<td>1</td>
<td>1</td>
<td>Not Rejected</td>
</tr>
<tr>
<td>HITS(h, freq)</td>
<td>1</td>
<td>1</td>
<td>Not Rejected</td>
</tr>
<tr>
<td>HITS(a, bin)</td>
<td>1</td>
<td>1</td>
<td>Not Rejected</td>
</tr>
<tr>
<td>HITS(a, freq)</td>
<td>1</td>
<td>1</td>
<td>Not Rejected</td>
</tr>
<tr>
<td>LSI+Dyn+PR(bin)\text{top}</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td>Rejected</td>
</tr>
<tr>
<td>LSI+Dyn+PR(bin)\text{bottom}</td>
<td>0.004</td>
<td>0</td>
<td>Rejected</td>
</tr>
<tr>
<td>LSI+Dyn+PR(freq)\text{top}</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td>Rejected</td>
</tr>
<tr>
<td>LSI+Dyn+PR(freq)\text{bottom}</td>
<td>&lt; 0.0001</td>
<td>0.74</td>
<td>Not Rejected</td>
</tr>
<tr>
<td>LSI+Dyn+HITS(a, freq)\text{top}</td>
<td>0</td>
<td>&lt; 0.0001</td>
<td>Rejected</td>
</tr>
<tr>
<td>LSI+Dyn+HITS(a, freq)\text{bottom}</td>
<td>&lt; 0.0001</td>
<td>0.99</td>
<td>Not Rejected</td>
</tr>
<tr>
<td>LSI+Dyn+HITS(h, freq)\text{top}</td>
<td>0</td>
<td>1</td>
<td>Not Rejected</td>
</tr>
<tr>
<td>LSI+Dyn+HITS(h, freq)\text{bottom}</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td>Rejected</td>
</tr>
<tr>
<td>LSI+Dyn+HITS(a, bin)\text{top}</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td>Rejected</td>
</tr>
<tr>
<td>LSI+Dyn+HITS(a, bin)\text{bottom}</td>
<td>&lt; 0.0001</td>
<td>1</td>
<td>Not Rejected</td>
</tr>
<tr>
<td>LSI+Dyn+HITS(h, bin)\text{top}</td>
<td>0</td>
<td>1</td>
<td>Not Rejected</td>
</tr>
<tr>
<td>LSI+Dyn+HITS(h, bin)\text{bottom}</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td>Rejected</td>
</tr>
</tbody>
</table>
Research Questions Revisited

- **RQ1**: Does combining web mining algorithms with an existing approach to feature location improve its effectiveness?
  - Yes

- **RQ2**: Which web-mining algorithms, HITS or PageRank, produces better results?
  - HITS
Best Techniques

- LSI+Dyn+HITS\(h, \text{ freq}\)\text{bottom}
- LSI+Dyn+HITS\(h, \text{ bin}\)\text{bottom}

- Methods with low HITS hub values are getters and setters
Current Work (not in the paper)

- HITS and PageRank on static vs. dynamic info
- Evaluation first relevant vs. all relevant methods
- Evaluation against fan-in and fan-out and heuristics based on setters and getters
- Impact of thresholds on the filtering power
Tool Support

- FLAT$^3$
  - Eclipse Plug-in
  - Lucene-based IR
  - Execution tracing
  - Integration
  - Tagging
  - Metrics

http://www.cs.wm.edu/semeru/flat3/

Summary

- Proposed and implemented novel methods for feature location based combinations of:
  - Textual analysis, dynamic analysis and web mining
- Evaluated proposed methods on large, open-source systems
- Developed practical tools for the proposed approaches
- Released benchmarks for feature location:
Thank you. Questions?

SEMERU @ William and Mary

http://www.cs.wm.edu/semeru/
denys@cs.wm.edu