Combining Formal Concept Analysis with Information Retrieval for Concept Location in Source Code

Denys Poshvyvanyk and Andrian Marcus

SEVERE Group @

ICPC 2007, Banff, Alberta
Incremental Change of Software

- Change request
- Concept Location
- Impact Analysis
- Implementation
- Change Propagation
- Testing

ICPC 2007, Banff, Alberta
State of the Art in Concept Location

- **Static**
  - Dependency based search [Rajlich’00]
  - Information Retrieval based methods [Marcus’04]

- **Dynamic**
  - Execution traces - Reconnaissance [Wilde’92]
  - Scenario based probabilistic ranking [Antoniol’06]

- **Combined**
  - FCA + Execution Traces [Eisenbarth’03][Tonella’04]
  - IR + Execution Traces [Poshyvanyk’07]
  - ...
### IRiSS View

<table>
<thead>
<tr>
<th>Problems</th>
<th>Javadoc</th>
<th>IRiSS View</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camera</td>
<td>setScreenParams</td>
<td>0.813055</td>
<td></td>
</tr>
<tr>
<td>ScaleObjectTool</td>
<td>mouseDragged</td>
<td>0.807305</td>
<td></td>
</tr>
<tr>
<td>Camera</td>
<td>setScreenParamsParallel</td>
<td>0.80328</td>
<td></td>
</tr>
<tr>
<td>Camera</td>
<td>setSize</td>
<td>0.788526</td>
<td></td>
</tr>
<tr>
<td>ViewerCanvas</td>
<td>setSize</td>
<td>0.723398</td>
<td></td>
</tr>
<tr>
<td>GLCanvasDrawer</td>
<td>prepareView3D</td>
<td>0.705305</td>
<td></td>
</tr>
<tr>
<td>ViewerCanvas</td>
<td>scaleChanged</td>
<td>0.700639</td>
<td></td>
</tr>
<tr>
<td>JitterModule</td>
<td>setZScale</td>
<td>0.690434</td>
<td></td>
</tr>
<tr>
<td>JitterModule</td>
<td>setXScale</td>
<td>0.690117</td>
<td></td>
</tr>
<tr>
<td>ImageModule</td>
<td>setXScale</td>
<td>0.690104</td>
<td></td>
</tr>
</tbody>
</table>

### GES Search

Enter the search string:

**Animation Preview**

Console:

GES search: Animation Preview in workspace (no JRE) - 250 matches found.

1. ActorEditorWindow.java - Art of Illusion/ArtOfIllusion/src/artofillusion/animation (11 matches)
2. AnimatorPreviewer.java - Art of Illusion/ArtOfIllusion/src/artofillusion/animation (19 matches)
3. CProgressObject.java - Art of Illusion/ArtOfIllusion/src/artofillusion/object (2 matches)
4. CustomDistortionTrack.java - Art of Illusion/ArtOfIllusion/src/artofillusion/animation/distortion (10 matches)
5. CylinderMapping.java - Art of Illusion/ArtOfIllusion/src/artofillusion/texture (12 matches)
6. LinearMapping3D.java - Art of Illusion/ArtOfIllusion/src/artofillusion/texture (12 matches)
7. ObjectInfo.java - Art of Illusion/ArtOfIllusion/src/artofillusion/object (13 matches)
8. ObjectTextureDialog.java - Art of Illusion/ArtOfIllusion/src/artofillusion (42 matches)
9. ProcedurePositionTrack.java - Art of Illusion/ArtOfIllusion/src/artofillusion/animation (9 matches)

### IRiSS - persistent cache

- CTextView::WrapLineCached: confidence = 0.87
- CTextView::InvalidateLineCache: confidence = 0.84
- CTextView::DrawSingleLine: confidence = 0.82
- CTextView::OnDraw: confidence = 0.79
- CTextView::SetFont: confidence = 0.68
- CTextView::OnSize: confidence = 0.59
- CTextView::CalculateActualOffset: confidence = 0.57

---

ICPC 2007, Banff, Alberta
ICPC 2007, Banff, Alberta
Clustered Results

Cluster Conference contains 7 documents. (Details)

Sponsored Results for conference

1. Discount Hotel Rates
   5 Room Minimum. Discount rates for conferences, hotels & meetings.
   www.Grouplo.com - Sponsored Listings 1

2. 6.5¢ Toll-Free Conference
   Fortune 500 service provider... Now available for you!
   SmartConferenceNow.com - Sponsored Listings 2

1. www.icpc4cops.org - INDEX HTML --> International Conference
   Information about the specialized pastoral care ministry provided by man and women who serve as police chaplains in law enforcement agencies.
   www.icpc4cops.org - Ask 4, MSN 14, Wisesnut 21

2. ICPC Home :: IEEE International Conference on Program Comprehension
   Program comprehension is a vital software engineering and maintenance activity. It is necessary to facilitate reuse, inspection, maintenance, reverse engineering, reengineering, migration ...
   www.program-comprehension.org - MSN 6, Wisesnut 17, Ask 44

3. ICPC 2007: Home
   15th IEEE International Conference on Program Comprehension June 26–29, 2007 - Banff, Alberta, Canada
   www.cs.ualberta.ca/icpc2007 - Ask 14, MSN 37

4. AVL Internet Conference Area ICPC - International Commercial Powertrain Conf...
   Company > Conference Area > ICPC - International Commercial Powertrain Conference Engine & Environment ICPC - International Commercial Powertrain Conference Large Engines TechDay AST User ...
   www.avl.com/.../encoded/YXByPWJbXMmcDFnZT12aWV3Jim5vZGVpZD00MDAwMTGxNDQ_3D.html

ICPC 2007, Banff, Alberta
Concept Location with Concept Lattices

1. Creating a corpus of a software system
2. Indexing with Latent Semantic Indexing
3. Formulating a query
4. Ranking methods
5. Selecting words
6. Clustering with Formal Concept Analysis
7. Examining results
Creating a corpus of a software system

- Parsing source code and extracting documents
  - corpus - collection of documents (e.g., methods)

- Removing non-literals and stop words
  - common words in English, standard function library names, programming language keywords

- Preprocessing: split_identifiers and SplitIdentifiers
Concept Location with Concept Lattices

1. Creating a corpus of a software system
2. Indexing with Latent Semantic Indexing
3. Formulating a query
4. Ranking methods
5. Selecting words
6. Clustering with Formal Concept Analysis
7. Examining results
Concept Location with Concept Lattices

1. Creating a corpus of a software system
2. Indexing with Latent Semantic Indexing
3. Formulating a query
4. Ranking methods
5. Selecting words
6. Clustering with Formal Concept Analysis
7. Examining results
Selecting Descriptive Words

• Ranking criteria - words which are mostly similar to search results than to the rest of the system
• Words in a subset of search results are ranked with LSI

<table>
<thead>
<tr>
<th>Method</th>
<th>Page</th>
<th>Paper</th>
<th>Rendering</th>
<th>Device</th>
<th>Printer</th>
</tr>
</thead>
<tbody>
<tr>
<td>startPage</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>endPage</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>getBounds</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>otherMethod1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>otherMethod2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
Concept Location with Concept Lattices

1. Creating a corpus of a software system
2. Indexing with Latent Semantic Semantic Indexing
3. Formulating a query
4. Ranking methods
5. Selecting words
6. Clustering with Formal Concept Analysis
7. Examining results
FCA Example - classification of geometrical shapes

- square
- equilateral-triangle
- isosceles-triangle
- scalene-triangle
- rectangle

Example from Paolo Tonella’s tutorial on Formal Concept Analysis in Software Engineering
FCA Example - classification of geometrical shapes

Example from Paolo Tonella’s tutorial on Formal Concept Analysis in Software Engineering
FCA Example - classification of geometrical shapes

Example from Paolo Tonella’s tutorial on Formal Concept Analysis in Software Engineering
FCA Example - classification of geometrical shapes

- square
- equilateral-triangle
- isosceles-triangle
- scalene-triangle
- rectangle
- regular

Example from Paolo Tonella’s tutorial on Formal Concept Analysis in Software Engineering
FCA Example - classification of geometrical shapes

Example from Paolo Tonella’s tutorial on Formal Concept Analysis in Software Engineering
**FCA Example - classification of geometrical shapes**

<table>
<thead>
<tr>
<th>Objects</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4-sides</td>
</tr>
<tr>
<td>square</td>
<td>¥</td>
</tr>
<tr>
<td>rectangle</td>
<td>¥</td>
</tr>
<tr>
<td>scalene-triangle</td>
<td></td>
</tr>
<tr>
<td>isoscele-triangle</td>
<td></td>
</tr>
<tr>
<td>equilateral-triangle</td>
<td></td>
</tr>
</tbody>
</table>

Example from Paolo Tonella’s tutorial on Formal Concept Analysis in Software Engineering
Concept lattice

Example from Paolo Tonella’s tutorial on Formal Concept Analysis in Software Engineering
Clustering Search Results with Formal Concept Analysis

- Subset of methods in search results
- Subset of words selected from search results
- Example: searching for *print page* feature

<table>
<thead>
<tr>
<th></th>
<th>printer</th>
<th>print</th>
<th>page</th>
<th>job</th>
<th>device</th>
<th>paper</th>
<th>rendering</th>
</tr>
</thead>
<tbody>
<tr>
<td>startJob</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>endJob</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cancelJob</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>startPage</td>
<td></td>
<td>×</td>
<td></td>
<td></td>
<td></td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>endPage</td>
<td></td>
<td></td>
<td>×</td>
<td></td>
<td></td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>getBounds</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>

ICPC 2007, Banff, Alberta
Concept Lattice of Search Results for Print Page

Attributes (words) - grey boxes
Objects (methods) - white boxes
Case Study

• Locating concepts associated with bug descriptions
  - Bug fixes contain a set of changed methods

• Source code of Eclipse 3.1
  - vocabulary of unique terms - 56,863¹
  - number of extracted methods - 86,208

¹Oxford English Dictionary has 171,476 words in current use
Evaluation of Novel Approach

- Lattices compared with ranked lists:
  - Number of methods: 80, 90, 100
  - Number of words: 10, 15, 20, 25

- Studied properties of lattices:
  - Grouping of relevant information
  - Browsing overhead of lattice structure

- Redefined measures from [Cigarran’04]:
  - Lattice distillation factor
  - Lattice browsing complexity
Measures

Lattice Distillation Factor
- how many methods are visited

Lattice Browsing Complexity
- how many concept nodes (categories) are visited while browsing

Ranked list
1. startPage
2. endPage
3. getBounds
4. startJob
5. endJob
6. cancelJob
...

ICPC 2007, Banff, Alberta
Locating Features in Eclipse - Example

- Table Headers Feature

- “The task list, which uses the native table widget, cannot be sorted by clicking on the table headers”

- Query: “table headers sorted”

- Associated with bug report# 34160
Locating *Table Headers* Feature

Clustered results into labeled categories

- **Table**
  - createTable
    - Widget.setData
    - FilteredList.TableUpdater
    - ...
    - Table.createWidget
  - tableViewer
  - getTable
  - tableValue, keyTable
- **Header**
  - setHeaderVisible
  - setLineVisible
  - ...

**Ranked List**

1. WidgetTable.put
2. TableTree.getTable
3. EditorsView.getTable
4. SimpleLookupTable.rehash
5. WidgetTable.shells
   - ...
39. TableTreeEditor.resize
   - ...
71. Widgets.Table.createWidget
Results for *Table Headers*

<table>
<thead>
<tr>
<th>Docs</th>
<th>Terms</th>
<th>$L_{\text{MBA}}$</th>
<th>C</th>
<th>$C_{\text{VIEW}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>10</td>
<td>51</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>100</td>
<td>15</td>
<td>40</td>
<td>25</td>
<td>11</td>
</tr>
<tr>
<td>100</td>
<td>20</td>
<td>38</td>
<td>33</td>
<td>11</td>
</tr>
<tr>
<td>100</td>
<td>25</td>
<td>36</td>
<td>39</td>
<td>10</td>
</tr>
<tr>
<td>90</td>
<td>10</td>
<td>43</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>90</td>
<td>15</td>
<td>36</td>
<td>24</td>
<td>11</td>
</tr>
<tr>
<td>90</td>
<td>20</td>
<td>34</td>
<td>31</td>
<td>11</td>
</tr>
<tr>
<td>90</td>
<td>25</td>
<td>26</td>
<td>36</td>
<td>11</td>
</tr>
<tr>
<td>80</td>
<td>10</td>
<td>38</td>
<td>15</td>
<td>7</td>
</tr>
<tr>
<td>80</td>
<td>15</td>
<td>30</td>
<td>22</td>
<td>10</td>
</tr>
<tr>
<td>80</td>
<td>20</td>
<td>29</td>
<td>28</td>
<td>10</td>
</tr>
<tr>
<td>80</td>
<td>25</td>
<td>23</td>
<td>33</td>
<td>10</td>
</tr>
</tbody>
</table>

- $L_{\text{MBA}}$ - rank of the first relevant method in the lattice
- $C$ - number of concepts in the lattice
- $C_{\text{VIEW}}$ - number of expanded concepts in the lattice
Conclusions

• Novel representation for the search results in the source code
  - results are ranked and \textit{structured}
  - labels for topics are \textit{automatically} extracted

• Scalable to software of \textit{any size}
  - fixed number of top search results
  - fixed number of attributes
Current and Future Work

• Different strategies for ranking and selecting words
• Lattice size: best ratio of docs and terms
• Adding IR ranks and similarities to the lattices
• Use FCA and LSI during impact analysis
• More user studies
• Search problem inside the search problem