Recovering Traceability Links in Software Artifact Management System using Information Retrieval Methods

Andrea De Lucia, Fausto Fasano, Rocco Oliveto, and Genoveffa Tortora

(TOSEM 2007)
Software artifact traceability

- Definition – linking between the artifacts that are dependable each other

Importance
- Giving essential support in understanding the relationships existing within and across software requirements, design, and implementation
Software artifact traceability (Cont’d)

- Problem
  - Support for traceability and tool is not satisfactory
    - Expensive activity of manual detection and maintenance

Research goal

- Provide a tool to support traceability recovery using Latent Semantic Method (LSI)
  - Discover emerging traceability links during software evolution and to monitor previously traced links
Latent Semantic Indexing (LSI)

- Documents represented by index terms
- Query by extracting information about the occurrences of terms within them
  - Vector Space Model (VSM)
    - Represent similarity between documents as cosine product of the vectors
  - Singular Value Decomposition (SVD)
    - Overcome the synonymy and polysemy problem, which occur with the VSM Model
Vector Space Model (VSM)

Document and queries are represented as vectors of terms that occur within documents

- Term-by-document matrix $A$
  - m by n matrix (m terms with n documents)

$$a_{i,j} = L(i, j) \cdot G(i)$$

where $1 < i < m$, $1 < j < n$

$a_{i,j}$: weight of the $i^{th}$ term in $j^{th}$ document
$L(i, j)$: frequency of the terms in the documents
$G(i)$: global weight decrease with the wide spread of the $i^{th}$ term in entire document space
Example of VSM

ReqSpecification
Register for Course
This use case allows a student to register for course offering process …

Code
CourseInfo
Public class CourseInfo{
    int Course_ID;
    ...
}

\[
\begin{bmatrix}
2.6 & 5.6 & \cdots \\
\vdots & \vdots & \vdots \\
1.0 & 1.5 & \cdots
\end{bmatrix}
\]

Cosine of angle between two vectors

\[
\cos(x) = 1 \text{ if } x = 0^\circ \\
0 \text{ else } x = 90^\circ
\]
Singular Value Decomposition

- Decompose the term-by-document matrix into the product of three other matrices
  - Overcome the synonymy and polysemy problem

\[ A = T S D \]

- \( T \): m by r matrix of the terms (orthogonal)
- \( S \): r by r diagonal matrix of singular values
- \( D \): r by n matrix of the documents (orthogonal)
Singular Value Decomposition (Cont’d)

- **S • D**
  - Cluster related terms with respect to documents
  - Reduce the space and time to compute the weight
    - 20% of numbers of original document is proper

- Cosine of angle between two vector is a measure of similarity
Assessing LSI (1/4)

- Adjust parameters to use LSI as traceability recovery tool (150 documents, 865 terms)
  - Categorize according to a type of the artifacts

<table>
<thead>
<tr>
<th>Artifact Category</th>
<th>Without categorization</th>
<th>With categorization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use cases</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interaction diagrams</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test cases</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Code classes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\text{recall} = \frac{\text{correct} \cap \text{retrieved}}{\text{correct}}
\]

\[
\text{precision} = \frac{\text{correct} \cap \text{retrieved}}{\text{retrieved}}
\]
Choose a method to cut the ranked lists

- Candidate link has the similarity value over threshold
  - Variable threshold – specify the minimum and maximum similarity values
  - Variable cut point – specify the percentage of the links
  - Scale threshold – percentage of the best similarity value between two artifacts

![Graph showing precision vs. recall with annotations for variable threshold, variable cut point, and scale threshold.](image-url)
### Assessing LSI(3/4)

**Assessing to extract the traceability links**

<table>
<thead>
<tr>
<th>Source category</th>
<th>Target category</th>
<th>Possible links</th>
<th>Correct links</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use case</td>
<td>Use case</td>
<td>870</td>
<td>86</td>
</tr>
<tr>
<td>Use case</td>
<td>Interaction Diagram</td>
<td>600</td>
<td>26</td>
</tr>
<tr>
<td>Use case</td>
<td>Test case</td>
<td>1890</td>
<td>63</td>
</tr>
<tr>
<td>Use case</td>
<td>Code class</td>
<td>1110</td>
<td>93</td>
</tr>
<tr>
<td>Interaction Diagram</td>
<td>Interaction Diagram</td>
<td>380</td>
<td>35</td>
</tr>
<tr>
<td>Interaction Diagram</td>
<td>Test case</td>
<td>1260</td>
<td>83</td>
</tr>
<tr>
<td>Interaction Diagram</td>
<td>Code class</td>
<td>740</td>
<td>69</td>
</tr>
<tr>
<td>Test case</td>
<td>Test case</td>
<td>3906</td>
<td>289</td>
</tr>
<tr>
<td>Test case</td>
<td>Code class</td>
<td>2331</td>
<td>204</td>
</tr>
<tr>
<td>Code class</td>
<td>Code class</td>
<td>1332</td>
<td>57</td>
</tr>
<tr>
<td><strong>Total number of correct links</strong></td>
<td></td>
<td><strong>1005</strong></td>
<td></td>
</tr>
</tbody>
</table>
Assessing LSI (4/4)

Accessing to extract the traceability links (Cont’d)

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.95</td>
<td>80%</td>
<td>30%</td>
</tr>
<tr>
<td>0.9</td>
<td>90%</td>
<td>40%</td>
</tr>
<tr>
<td>0.85</td>
<td>95%</td>
<td>45%</td>
</tr>
<tr>
<td>0.8</td>
<td>100%</td>
<td>50%</td>
</tr>
<tr>
<td>0.75</td>
<td>90%</td>
<td>55%</td>
</tr>
<tr>
<td>0.7</td>
<td>75%</td>
<td>60%</td>
</tr>
<tr>
<td>0.65</td>
<td>60%</td>
<td>65%</td>
</tr>
<tr>
<td>0.6</td>
<td>45%</td>
<td>70%</td>
</tr>
<tr>
<td>0.55</td>
<td>30%</td>
<td>75%</td>
</tr>
<tr>
<td>0.5</td>
<td>15%</td>
<td>80%</td>
</tr>
<tr>
<td>0.45</td>
<td>5%</td>
<td>85%</td>
</tr>
<tr>
<td>0.4</td>
<td>0%</td>
<td>90%</td>
</tr>
<tr>
<td>0.35</td>
<td>-</td>
<td>95%</td>
</tr>
<tr>
<td>0.3</td>
<td>-</td>
<td>100%</td>
</tr>
</tbody>
</table>

Recall – over 70%
Precision – over 30%

Reasonable to use LSI as traceability recovery method
Advanced Artifact Management System (ADAMS)

- Artifact-based process support system
- Latent Semantic Indexing (LSI) method is used
  - Useful to the identification of traceability links
  - Helpful in the identification of quality problems in the text description of software artifacts

- Modules
  - Indexer, LSI, Query
Indexer
- Build a hash table containing the number of occurrences of the different terms in the artifact
  - When an artifact is checked-in

LSI (Latent Semantic Indexing)
- Compute VSM and SVD periodically and whenever an artifact is checked-in

Query
- Software engineer chooses the subsets of source and target artifacts for traceability recovery
Functionality

- Software engineer combines manual tracing activity with the tool support
  - retrieved – the set of links retrieved by the tool
  - traced – the set of links traced by the software engineer

- As # of artifacts increase, S/W engineer can not trace all the correct link
Functionality (Cont’d)

- **Suggested link**
  - Trace link – tool is right and link should be traced
  - False positive – tool is wrong and it should be discarded

- **Warning link**
  - Trace link (False negative) – tool is wrong and link should be traced
    - It can be due to a poor description of the artifacts or quality problem of the tool
  - Discard link – tool is right and it should be discarded

- **Discarded links** is considered at next iterations only when the similarity of their links has increased
Experience and Evaluation (1/2)

- Environment of experiment

  - 17 projects with 150 students

<table>
<thead>
<tr>
<th></th>
<th>1st period</th>
<th>2nd period</th>
<th>3rd period</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># of Artifact</td>
<td># of Artifact</td>
<td>Added Artifact</td>
</tr>
<tr>
<td>RAD</td>
<td>760</td>
<td>1122</td>
<td>431</td>
</tr>
<tr>
<td>Code</td>
<td>23</td>
<td>39</td>
<td>23</td>
</tr>
</tbody>
</table>

*RAD: Requirement Analysis Document (requirement, scenarios, use cases, sequence diagrams)

- 2 TA – project manager, quality manager
## Analysis of Suggested links

- **Emerging link** – # of suggested links that increased their similarity as a consequence of changes on the documents
- **New link** – # of suggested links arising as a consequence of new artifacts

<table>
<thead>
<tr>
<th></th>
<th>1(^{\text{st}}) period</th>
<th>2(^{\text{nd}}) period</th>
<th>3(^{\text{rd}}) period</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MT (L)</td>
<td>Suggested links</td>
<td>MT (L)</td>
</tr>
<tr>
<td></td>
<td>TL</td>
<td>FP</td>
<td>TL</td>
</tr>
<tr>
<td><strong>RAD vs RAD</strong></td>
<td>275</td>
<td>441</td>
<td>1108</td>
</tr>
<tr>
<td><strong>UC vs Code</strong></td>
<td>0</td>
<td>62</td>
<td>187</td>
</tr>
</tbody>
</table>

M TL: Manually Traced Links
TL: Traced Links
FP: False Positives

*RAD: Requirement Analysis Document
Helpful to derive the trace link during iteration
Conclusion

❖ Contribution

- Provides an assessment of LSI as a traceability recovery technique
- Develop a tool to support a traceability recovery
  - Discover emerging traceability links during software evolution and to monitor previously traced links
  - Help software engineers to improve their recovery performance
    - Based on the correct link not traced by tool, the quality managers ask for improving a poor description of the artifacts
Discussion

❖ Critique
  ▪ Lack of the experiment related to code artifacts
  ▪ LSI depends on the quality of the description

❖ Merit
  ▪ Experiment for parameters is well performed
    • Various parameters for LSI method is decided through the experiments
Vector Space Model (VSM)

- Weight of the element in a matrix

\[ L(i,j) = \log (tf_{ij} + 1) \]

\[ G(i) = 1 \sum_{j=1}^{n} \frac{p_{ij} \log (p_{ij})}{\log(n)} \]

\[ p_{ij} = \frac{tf_{ij}}{\sum_{k=1}^{n} tf_{ik}} \]

It increases as much as the \( i^{th} \) term is spread across the documents.

Ex) a, the, is and etc

\( tf_{ij} \): frequency of the \( i^{th} \) term in \( j^{th} \) document
\( p_{ij} \): sum of frequency of the \( i^{th} \) term in all document
### Source Artefacts (5)

1. Patient management - Requirement (DRAFT)
2. Medicine management - Requirement (DRAFT)
3. Doctor management - Requirement (DRAFT)
4. Visit management - Requirement (DRAFT)
5. Management reports - Requirement (DRAFT)

### Target Artefacts (30)

1. Reserve visit - Use Case (DRAFT)
2. Insert medicine - Use Case (DRAFT)
3. Modify medicine - Use Case (DRAFT)
4. Delete medicine - Use Case (DRAFT)
5. Show medicines - Use Case (DRAFT)
6. Dispense medicine - Use Case (DRAFT)
7. Medicine restock - Use Case (DRAFT)
8. Delete patient - Use Case (DRAFT)
9. Show patients - Use Case (DRAFT)
10. Insert doctor - Use Case (DRAFT)
11. Modify doctor - Use Case (DRAFT)
12. Delete doctor - Use Case (DRAFT)
13. Show doctors - Use Case (DRAFT)
14. Allocate doctor - Use Case (DRAFT)
15. Schedule examination - Use Case (DRAFT)
16. Schedule day hospital - Use Case (DRAFT)
17. Schedule hospitalization - Use Case (DRAFT)
18. Examination Report - Use Case (DRAFT)
### Recovering Functional Requirement onto Use case traceability links

#### Threshold: 90%

#### Previous iteration statistics:
- Threshold: 95%
- Suggested links: 3
- Traced Links: 2
- False positives: 1

### Suggested Links

<table>
<thead>
<tr>
<th>ID</th>
<th>Source Artefact</th>
<th>ID</th>
<th>Target Artefact</th>
<th>Similarity measure</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>169</td>
<td>Doctor management (Requirement)</td>
<td>212</td>
<td>Delete doctor (Use Case)</td>
<td>94.71</td>
<td><img src="" alt=" " /></td>
</tr>
<tr>
<td>168</td>
<td>Medicine management (Requirement)</td>
<td>197</td>
<td>Modify medicine (Use Case)</td>
<td>94.27</td>
<td><img src="" alt=" " /></td>
</tr>
<tr>
<td>167</td>
<td>Patient management (Requirement)</td>
<td>209</td>
<td>Show patient (Use Case)</td>
<td>92.96</td>
<td><img src="" alt=" " /></td>
</tr>
<tr>
<td>168</td>
<td>Medicine management (Requirement)</td>
<td>199</td>
<td>Show medicine (Use Case)</td>
<td>92.96</td>
<td><img src="" alt=" " /></td>
</tr>
<tr>
<td>169</td>
<td>Doctor management (Requirement)</td>
<td>213</td>
<td>Show doctors (Use Case)</td>
<td>91.73</td>
<td><img src="" alt=" " /></td>
</tr>
<tr>
<td>169</td>
<td>Doctor management (Requirement)</td>
<td>207</td>
<td>Modify patient (Use Case)</td>
<td>90.62</td>
<td><img src="" alt=" " /></td>
</tr>
</tbody>
</table>

### False Positives

<table>
<thead>
<tr>
<th>ID</th>
<th>Source Artefact</th>
<th>ID</th>
<th>Target Artefact</th>
<th>Similarity measure</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>167</td>
<td>Patient management (Requirement)</td>
<td>211</td>
<td>Modify doctor (Use Case)</td>
<td>95.41</td>
<td><img src="" alt=" " /></td>
</tr>
</tbody>
</table>

Submit actions