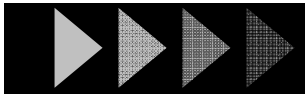


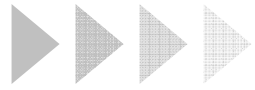
Recovering Traceability Links in Software Artifact Management System using Information Retrieval Methods

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

(TOSEM 2007)



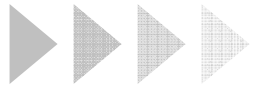
2007. 12. 5
HyeonJeong Kim



Contents



- 1. Introduction**
- 2. Background**
- 3. Assessing LSI**
- 4. ADAMS**
- 5. Experimental Result**
- 6. Conclusion**
- 7. Discussion**

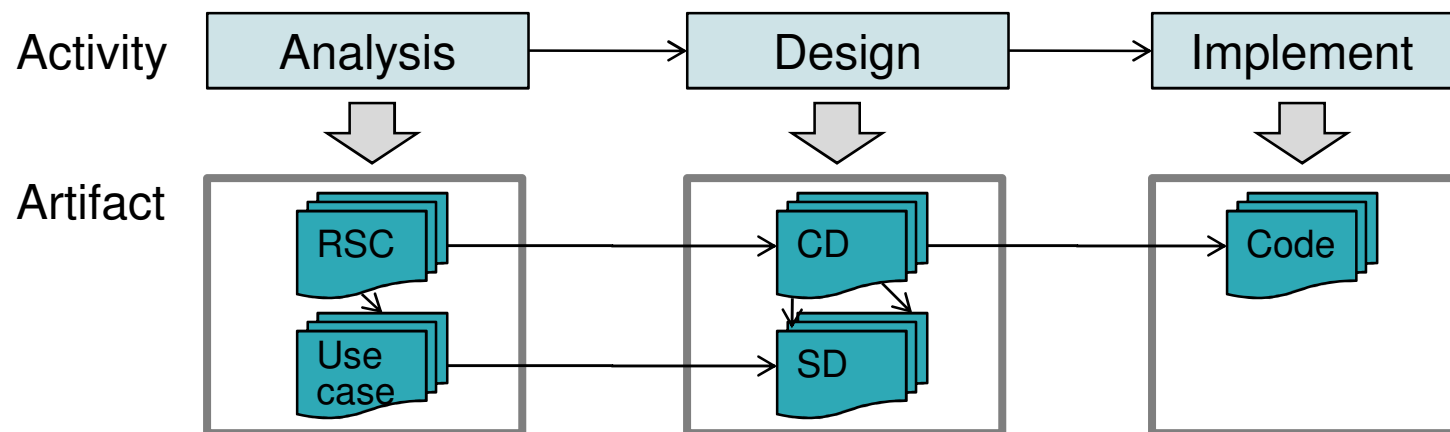


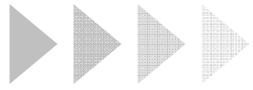
Introduction(1/2)



❖ 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

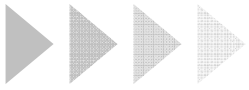




Introduction(2/2)



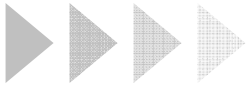
- ❖ Software artifact traceability(Cont'd)
 - Problem
 - Support for traceability and tool is not satisfactory
 - Expensive activity of manual detection and maintenance
- ❖ Research goal
 - Provide a tool to support traceability recovery using Latent Semantic Method(LSI)
 - Discover emerging traceability links during software evolution and to monitor previously traced links



Background(1/5)



- ❖ 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



Background(2/5)



❖ 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)

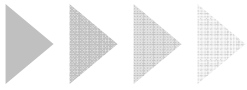
$$a_{i,j} = L(i, j) \cdot G(i)$$

$$\text{where } 1 < i < m, 1 < j < n$$

$a_{i,j}$: weight of the i^{th} term in j^{th} document

$L(i, j)$: frequency of the terms in the documents

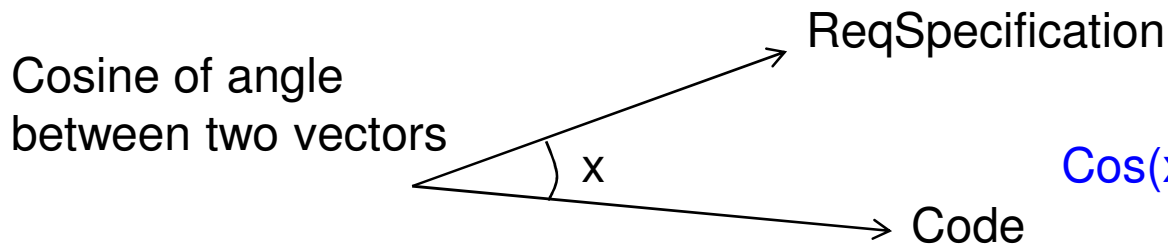
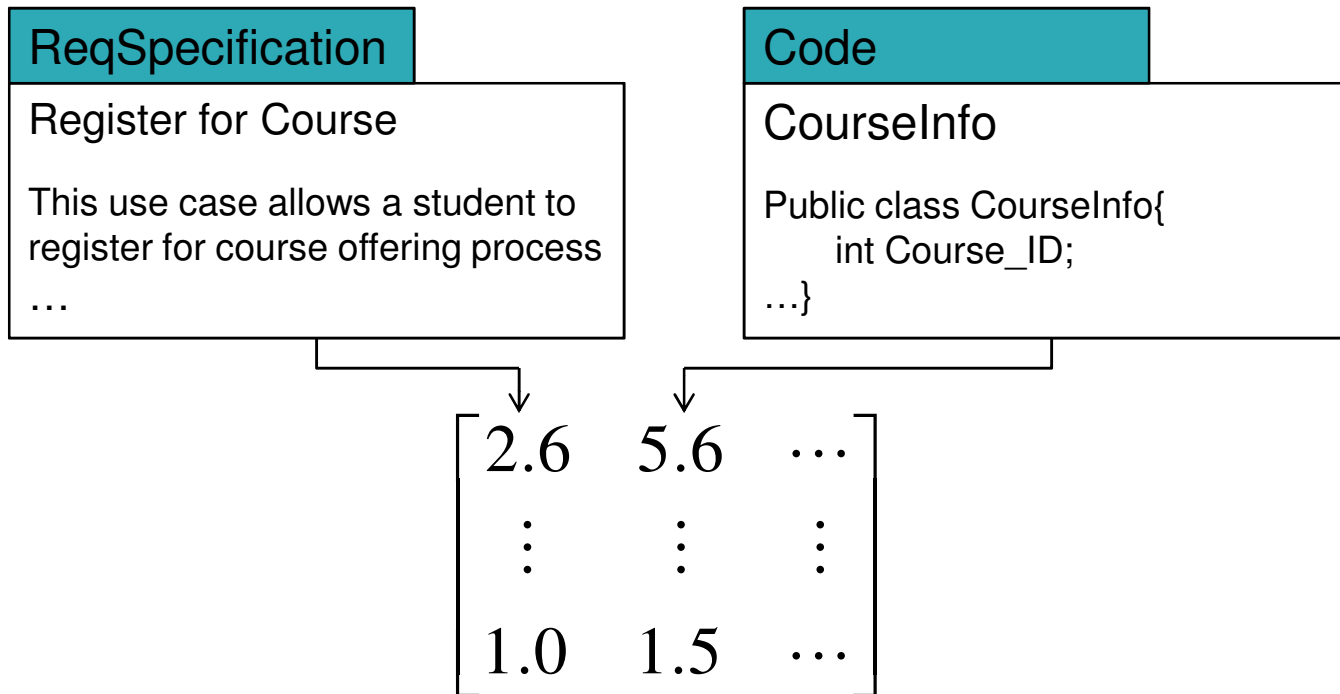
$G(i)$: global weight decrease with the wide spread of the i^{th} term in entire document space



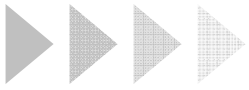
Background(3/5)



❖ Example of VSM



$$\begin{aligned} \text{Cos}(x) &= 1 \text{ if } x = 0^\circ \\ &0 \text{ else } x = 90^\circ \end{aligned}$$

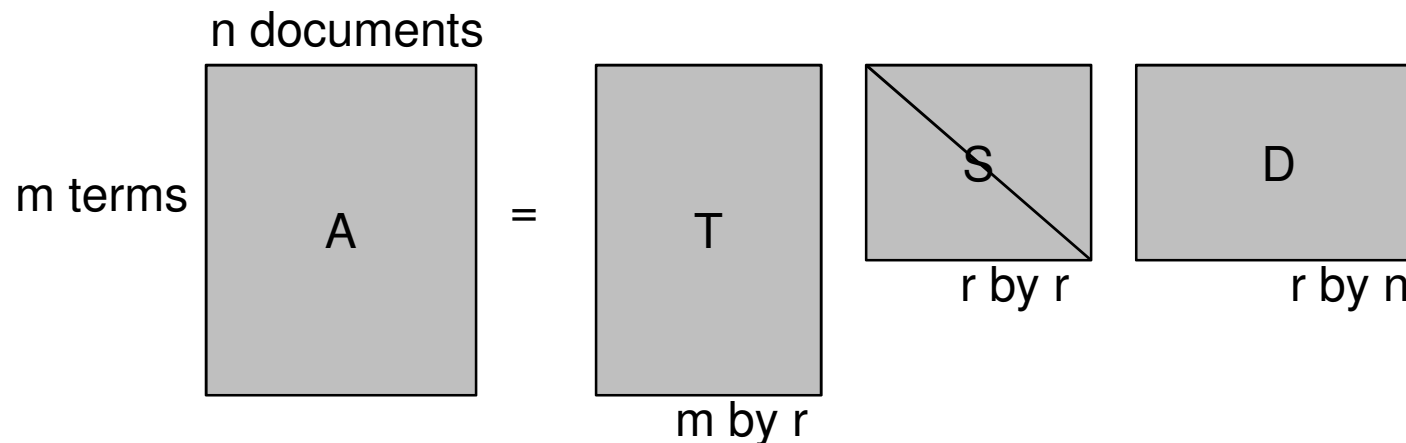


Background(4/5)

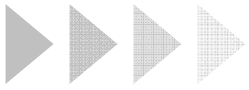


❖ Singular Value Decomposition

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



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)



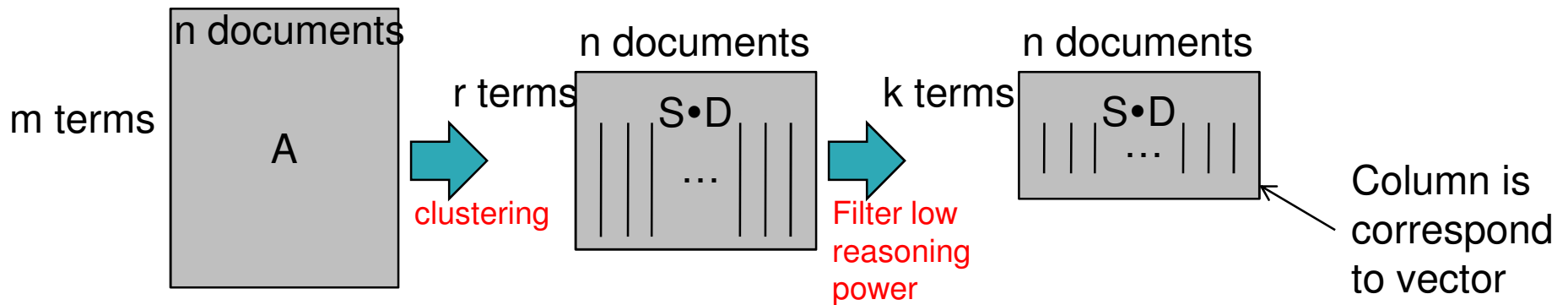
Background(5/5)



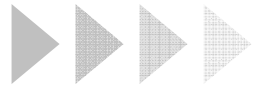
❖ Singular Value Decomposition(Cont'd)

■ $S \cdot D$

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



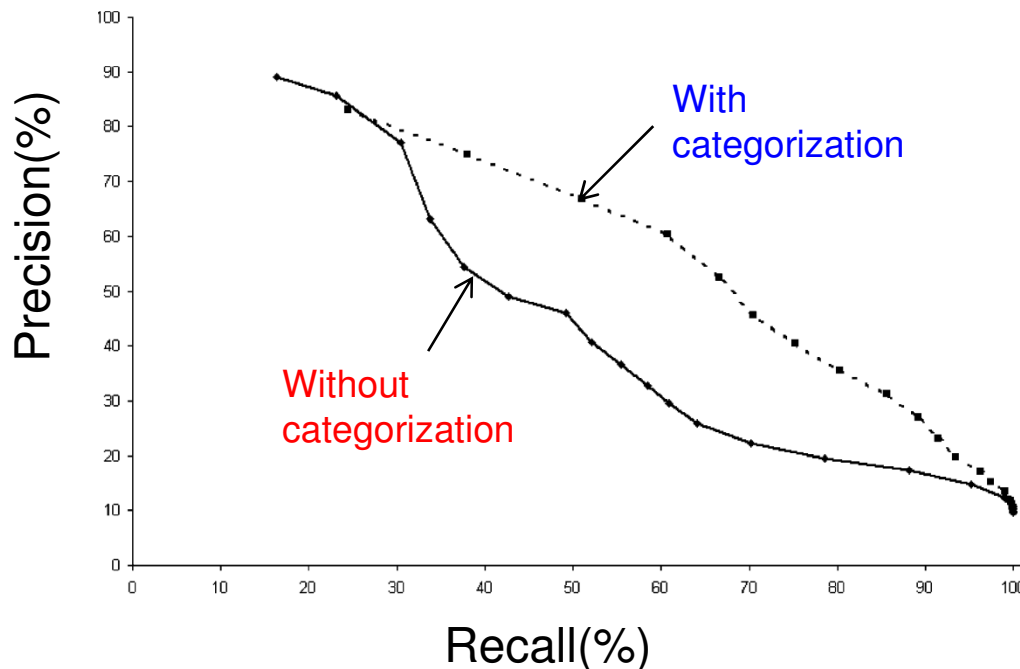
- Cosine of angle between two vector is a measure of similarity



Assessing LSI(1/4)



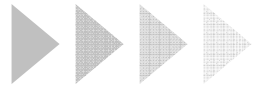
- ❖ Adjust parameters to use LSI as traceability recovery tool(150 documents, 865 terms)
 - Categorize according to a type of the artifacts



| Artifact Category |
|----------------------|
| Use cases |
| Interaction diagrams |
| Test cases |
| Code classes |
| Total number |

$$recall = \frac{\text{correct} \cap \text{retrieved}}{\text{correct}}$$

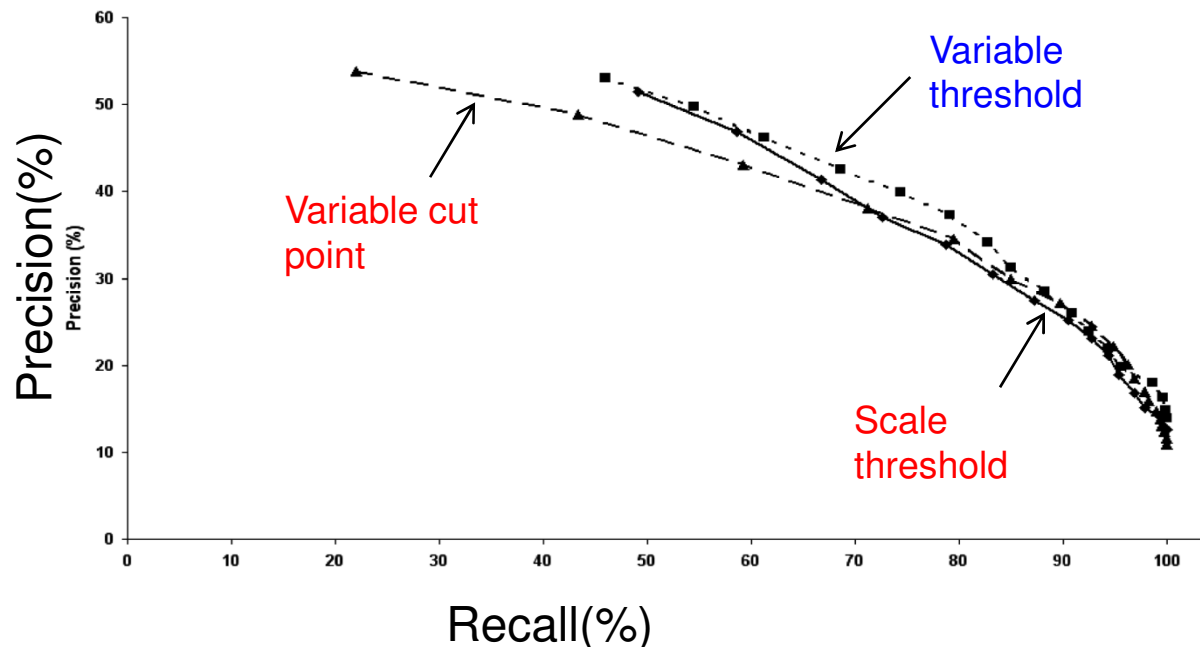
$$precision = \frac{\text{correct} \cap \text{retrieved}}{\text{retrieved}}$$

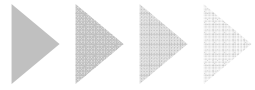


Assessing LSI(2/4)



- ❖ 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



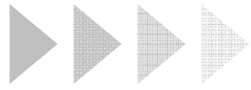


Assessing LSI(3/4)



❖ Assessing to extract the traceability links

| Source category | Target category | Possible links | Correct links |
|-------------------------------|---------------------|----------------|---------------|
| Use case | Use case | 870 | 86 |
| Use case | Interaction Diagram | 600 | 26 |
| Use case | Test case | 1890 | 63 |
| Use case | Code class | 1110 | 93 |
| Interaction Diagram | Interaction Diagram | 380 | 35 |
| Interaction Diagram | Test case | 1260 | 83 |
| Interaction Diagram | Code class | 740 | 69 |
| Test case | Test case | 3906 | 289 |
| Test case | Code class | 2331 | 204 |
| Code class | Code class | 1332 | 57 |
| Total number of correct links | | | 1005 |

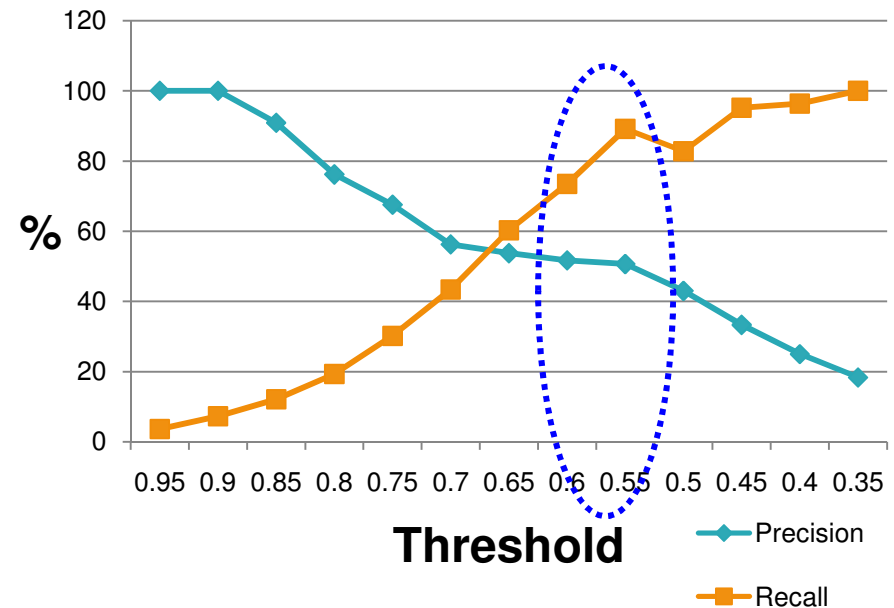
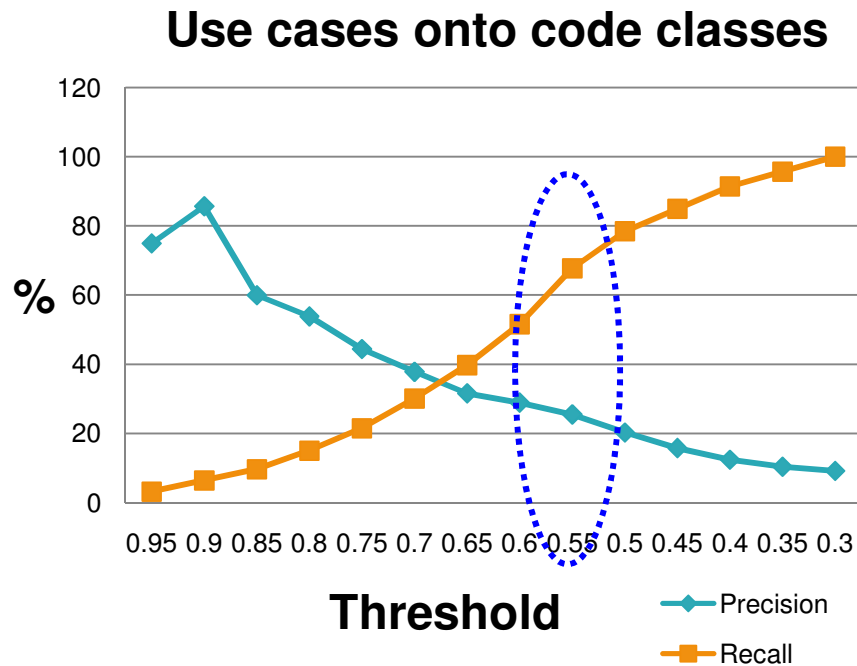


Assessing LSI(4/4)

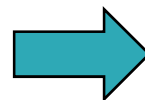


❖ Accessing to extract the traceability links(Cont'd)

Interaction diagrams onto test cases



Recall – over 70%
Precision – over 30%



Reasonable to use LSI as traceability recovery method

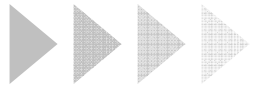


ADAMS(1/4)



- ❖ 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





ADAMS(2/4)



❖ Indexer

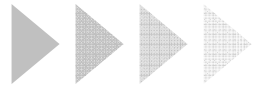
- Build a hash table containing the number of occurrences of the different terms in the artifact
 - When an artifact is checked-in

❖ 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

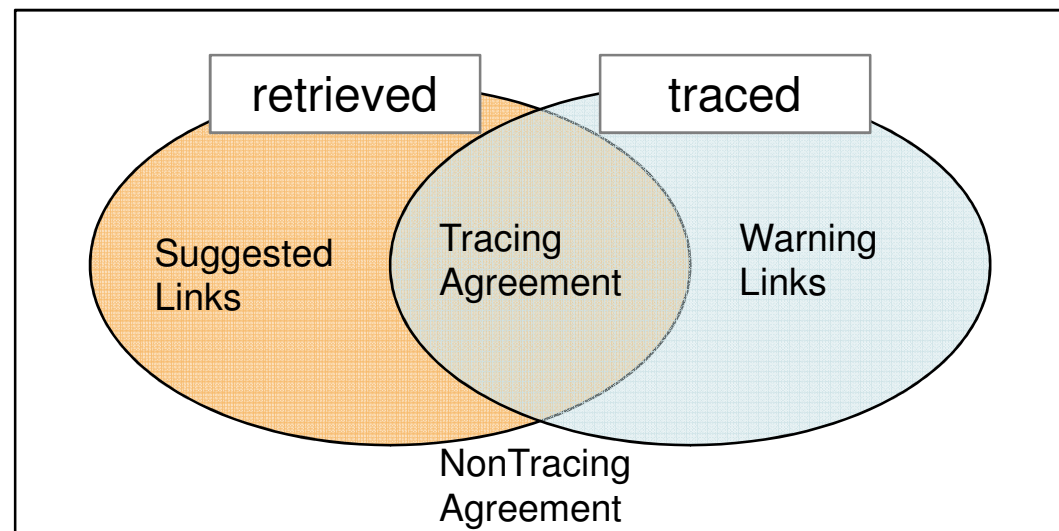


ADAMS(3/4)

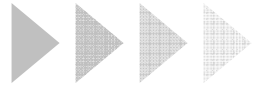


❖ Functionality

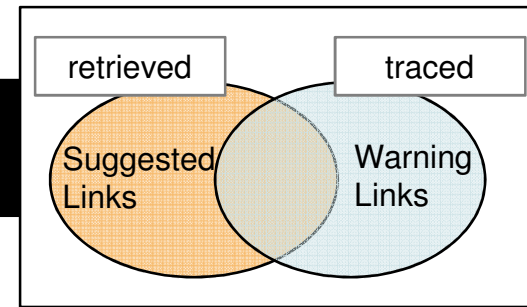
- Software engineer combines manual tracing activity with the tool support
 - retrieved – the set of links retrieved by the tool
 - traced – the set of links traced by the software engineer



- As # of artifacts increase, S/W engineer can not trace all the correct link



ADAMS(4/4)



❖ 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

| | 1 st period | 2 nd period | | 3 rd period | | | |
|------|------------------------|------------------------|----------------|------------------------|---------------|----------------|-------------------|
| | # of Artifact | # of Artifact | Added Artifact | Modified Artifact | # of Artifact | Added Artifact | Modified Artifact |
| RAD | 760 | 1122 | 431 | 407 | 1198 | 161 | 674 |
| Code | 23 | 39 | 23 | 11 | 156 | 117 | 21 |

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

- 2 TA – project manager, quality manager

Experience and Evaluation(2/2)

Suggested Links

retrieved

MTL

❖ 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

| | 1 st period | | | 2 nd period | | | | 3 rd period | | | | | |
|------------|------------------------|-----------------|------|------------------------|-----------------|-----|-----|------------------------|-----|-----------------|-----|-----|-----|
| | MTL | Suggested links | | MTL | Suggested links | | | | MTL | Suggested links | | | |
| | | | | | Emerge | | New | | | Emerge | | New | |
| | | TL | FP | | TL | FP | TL | FP | | TL | FP | TL | FP |
| RAD vs RAD | 275 | 441 | 1108 | 296 | 48 | 184 | 447 | 1207 | 53 | 36 | 255 | 121 | 597 |
| UC vs Code | 0 | 62 | 187 | 10 | 15 | 15 | 66 | 130 | 10 | 8 | 12 | 200 | 333 |

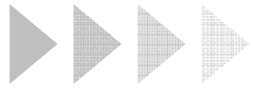
*RAD: Requirement Analysis Document

MTL: Manually Traced Links

Helpful to derive the trace link during iteration

TL: Traced Links

FP: False Positives

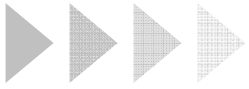


Conclusion



❖ 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

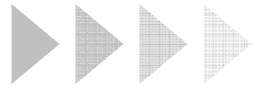


❖ Critique

- 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

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

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

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

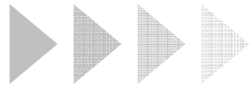


It increases as much as the i^{th} term is spread across the documents

Ex) a, the, is and etc

$tf_{i,j}$: frequency of of the i^{th} term in j^{th} document

$p_{i,j}$:sum of frequency of of the i^{th} term in all document



ADAMS

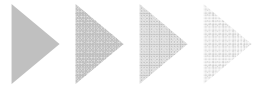


AD.A.M.S.

ADvanced Artefact Management System

User: Rocco Oliveto

| | | | | | |
|---|---|---|---|---|---|
| General Logout Refresh Home Administration Search <hr/> ToDo List All Projects My Projects My Artefacts <hr/> Project Artefacts Traceability Links Project Card | | SOURCE | | TARGET | |
| | Artefact Types | | <ul style="list-style-type: none"> Data Dictionary Item Deployment Diagram Design Goal Document Section Functional Requirement JSP Management Report Meeting Agenda Meeting Minute Node | | <ul style="list-style-type: none"> Statechart Diagram System Decomposition Test Case Test Execution Document Test Incident Test Plan UC Diagram UI Navigational Path UI Screen Mockup Use Case |
| | Artefact Filter | Contains <input type="button" value="v"/> | <input type="text" value="management"/> | Contains <input type="button" value="v"/> | <input type="text"/> |
| | <input type="button" value="Apply"/> | | | | |
| | | Source Artefacts (5) | | Target Artefacts (30) | |
| | <ul style="list-style-type: none"> 167) Patient management - Requirement (DRAFT) 168) Medicine management - Requirement (DRAFT) 169) Doctor management - Requirement (DRAFT) 170) Visit management - Requirement (DRAFT) 173) Management reports - Requirement (DRAFT) | | <ul style="list-style-type: none"> 193) Reserve visit - Use Case (DRAFT) 196) Insert medicine - Use Case (DRAFT) 197) Modify medicine - Use Case (DRAFT) 198) Delete medicine - Use Case (DRAFT) 199) Show medicines - Use Case (DRAFT) 200) Dispense medicine - Use Case (DRAFT) 201) Medicine restock - Use Case (DRAFT) 206) Insert patient - Use Case (DRAFT) 207) Modify patient - Use Case (DRAFT) 208) Delete patient - Use Case (DRAFT) 209) Show patients - Use Case (DRAFT) 210) Insert Doctor - Use Case (DRAFT) 211) Modify doctor - Use Case (DRAFT) 212) Delete doctor - Use Case (DRAFT) 213) Show doctors - Use Case (DRAFT) 214) Allocate doctor - Use Case (DRAFT) 220) Schedule examination - Use Case (DRAFT) 221) Schedule day hospital - Use Case (DRAFT) 222) Schedule hospitalization - Use Case (DRAFT) 225) Examination Report - Use Case (DRAFT) | | |
| | <input type="button" value="Select all"/> <input type="button" value="Invert selection"/> | | <input type="button" value="Select all"/> <input type="button" value="Invert selection"/> | | |
| <input type="button" value="Confirm"/> | | | | | |



- General
- Logout
- Refresh
- Home
- Administration
- Search

- ToDo List
- All Projects
- My Projects
- My Artefacts

- Project
- Artefacts
- Traceability Links
- Project Card

- Link analysis
- New analysis

Recovering Functional Requirement onto Use case traceability links

Threshold: %

Previous iteration statistics:

| | |
|-----------------|-----|
| Threshold | 95% |
| Suggested links | 3 |
| Traced Links | 2 |
| False positives | 1 |

| Suggested Links | | | | | | |
|-----------------|-----------------------------------|-----|----------------------------|--------------------|-----------------------|----------------------------|
| ID | Source Artefact | ID | Target Artefact | Similarity measure | Action | |
| | | | | | Trace link | Classify as False Positive |
| 169 | Doctor management (Requirement) | 212 | Delete doctor (Use Case) | 94.71 | <input type="radio"/> | <input type="radio"/> |
| 168 | Medicine management (Requirement) | 197 | Modify medicine (Use Case) | 94.27 | <input type="radio"/> | <input type="radio"/> |
| 167 | Patient management (Requirement) | 209 | Show patient (Use Case) | 92.66 | <input type="radio"/> | <input type="radio"/> |
| 168 | Medicine management (Requirement) | 199 | Show medicine (Use Case) | 92.66 | <input type="radio"/> | <input type="radio"/> |
| 169 | Doctor management (Requirement) | 213 | Show doctors (Use Case) | 91.73 | <input type="radio"/> | <input type="radio"/> |
| 169 | Doctor management (Requirement) | 207 | Modify patient(Use Case) | 90.62 | <input type="radio"/> | <input type="radio"/> |

| False Positives | | | | | |
|-----------------|----------------------------------|-----|--------------------------|--------------------|-----------------------|
| ID | Source Artefact | ID | Target Artefact | Similarity measure | Action |
| | | | | | Trace link |
| 167 | Patient management (Requirement) | 211 | Modify doctor (Use Case) | 95.41 | <input type="radio"/> |

Submit actions