Automated Linking of Natural Language Software Artifacts

- A Research Overview

Markus Borg
Software Engineering Research Group
Dept. of Computer Science
About me

• PhD Student
  – Software Engineering Research Group, since Jan 2010
• Three years at ABB, Malmö
  – Process Automation
• Research interests
  – Requirements-Test Alignment
  – Traceability
Outline of the Presentation

• Context and Motivation
• Information Retrieval (IR)
• IR-based Traceability Recovery
• A Research Overview
Context and Motivation
Large-scale Development

• Industrial development generated lots of information
  – Must be able to navigate this info space!
Traceability

• The ability to describe and follow the artifact life-cycle
  – Example: a use case is implemented by one or more classes that are tested by a set of test cases
Why Traceability?

• It is required or suggested by many standards
  – MIL-STD-498, IEEE/EIA 12207 (Military)
  – ISO/IEC 12207
  – DO178B, DO254 (Avionic)
  – EN50128 (Railways)

• Understand your system

• Change impact analysis

• Requirements traceability (forwards/backwards)
  – Good source for metrics

• Support software reuse
Traceability Can Exist Between

- Requirements and source code
- Source code and design
- Requirement and test cases
- Design and requirements
- Bug report and manual page
- Manual page to requirements
- ...
Traceability Challenges

• Maintaining traceability links during software evolution
  – Endless and error prone task
  – Information not updated or it is not there at all
  – Poor traceability contributes to project delays and failures

• State-of-the-practice tools do not provide sufficiently good support for traceability link generation and maintenance
  – Manually managed traceability matrix
Information Retrieval (IR)
What is Information Retrieval?

• The process of actively seeking out information relevant to a topic of interest
  (van Rijsbergen)

• Document
  – Generic term for an information holder (book, test case description, article, wiki page, source code file, method, requirement, etc.)

• Basis for Internet search engines
Why Analyze Textual Information in SE?

• Text is the common form of information representation
  – different abstraction levels
  – time dimension
• No predefined structure, grammar, vocabulary required
• Can be applied on legacy systems
Textual Content Everywhere

Pre-study

SRS

Func. spec.

System design

Detailed design

Requirements

Functional specification

Architecture design

Module design

Coding

Source code

Integration testing

Unit testing

Test scripts

Func. test spec.

System test spec.

Test report

Test report

System test spec.

Acceptance testing

System testing

Test report
Textual Content Everywhere

• And also…

- Bug reports
- Web pages
- Guides
- User manuals
- Marketing
- Emails
- Market analysis
- Dev. tools
- Sales summary
- Wikis
- Build logs
- Safety regulations
- Crib
- Todo lists
- Notes
- Safety regulations
IR-based Traceability Recovery
Main Idea and Assumption

• Artifacts having a high textual similarity are good candidates to establish links between
• Developers use consistent naming in various artifacts

• Challenges
  – Synonyms
  – Testers, developers, support engineers etc.
IR-based Traceability Tools

• Search once per document
  – Every document is used once as search query in the total document space

• Replace the degree of similarity between two documents with the probability of existence of a traceability link
### Details

**Id**: SRS41710  
**Type**: Functional  
**Section**: Service: Maintenance

**Description**: The operator shall be given an error tone if trying to remove a subscriber using a subscriber number and a multiple number that are not associated together.

---

### Details

**Id**: SRS42509  
**Type**: Functional  
**Section**: Service requirements - Maintenance

**Description**: If an operator tries to associate more than one subscriber number to a certain multiple number, an error tone shall be used.
Evaluating Tools

• Recall

\[
\frac{\text{# Relevant documents retrieved}}{\text{# Relevant documents}}
\]

• Precision

\[
\frac{\text{# Relevant documents retrieved}}{\text{# Retrieved documents}}
\]
A Research Overview
Publication Statistics

Traceability Recovery Publications

IR-Based Traceability Emp. Studies

(Web of Science)

(Intermediate results, Mapping study)
Tool Overview, Research Prototypes

- Poirot
- DrTrace
- ReqSimile
- TraceTool
- TraceWiz
- TRASE
- RETRO
- ReqAnalyst
- ADAMS Re-Trace

Dept. of Computer Science / Software Engineering Research Group
Case study, Antoniol et al. 2002

- LEDA 3.4
  - C++ abstract types library
  - 208 classes - 95 KLOC
  - Language: English
  - 88 manual pages
  - Traceability matrix reconstructed by hand
    - Manual pages to code: 124 correct links
Case study, De Lucia et al. 2008

- EasyClinic
  - Developed by final year master students
  - Language: Italian
  - 30 use cases, 20 sequence diagrams, 63 test cases, and 37 code classes
  - Traceability matrix provided by the original developers: 1005 correct links
Case study, De Lucia et al. 2008

• Lessons learned
  – The tool reduces the time spent by the software engineer
  – In general, the tool reduces tracing errors
  – Ability and Experience are influencing factors
  – The tool helps to reduce the gap between high and low ability subjects
  – The performances of the IR method is an influencing factor
Other Applications in Software Engineering

• Concept location (early step of impact analysis)
• Structural analysis
  – Coupling / Cohesion
  – Detection of code clones
• Bug triage
  – Duplicate detection
  – Developer recommendation
  – Automatic severity assignment
• And more…
Current Research Status

• Promising results
• Still few case studies on real industrial data, much focus on traceability to source code
  – Easier to get access to
• Stubborn hunt for recall and precision
  – Less focus on Return on Investment etc.
• The field matures
  – Recent publications from Daimler, IBM, Microsoft
Textual Content Everywhere

What is Information Retrieval?

- The process of actively seeking out information relevant to a topic of interest
  (van Riisbergen)
- Typically it refers to the automatic (rather than manual) retrieval of documents
- "Document" is the generic term for an information holder
  (e.g., chapter, article, webpage, class body, method, recruitment page, etc.)
- Based on Internet search engines

Tool Overview, Research Prototypes

LEDA (Antoniol et al. 2002)