A TraceLab-based Solution for Creating, Conducting, and Sharing Feature Location Experiments

Bogdan Dit, Evan Moritz, Denys Poshvyvanyk

20th IEEE International Conference on Program Comprehension (ICPC’12) - Passau, Germany
Have you ever tried to reproduce a case study and you couldn't?
Case Study in SE

• In general, it is difficult to reproduce case studies
  - Some details may be missing

• Results of case studies produce contradictory results *

In Feature Location (FL)...

- Case studies are hard to reproduce
  - Different datasets
  - Different implementations
  - Different metrics

Different Datasets

• Out of 60 papers... 5%
Different Datasets

• Out of 60 papers...

5%

95%

Different software systems
Different versions of the systems
Different data points from the same system
Different Implementation

- Often source code of FLTs is not available
- Configuration parameters are missing
Different Metrics

- Precision
- Recall
- F-measure
- Effectiveness measure
Other Observations

• Out of 60 papers...

Compared their technique against a limited number of existing techniques
Other Observations

- Out of 60 papers...

- 62% No Comparison
- 38% Compared their technique against a limited number of existing techniques
Why is Reproducibility of Case Studies Important?
Why is Reproducibility of Case Studies Important?

- Confirm if field is making progress
Why is Reproducibility of Case Studies Important?

• Confirm if field is making progress

• (Many requests for tools and data at ICPC’12)
Why is Reproducibility of Case Studies Important?

- Results can be checked by reviewers
- Results can be verified by other researchers
- Clarify some aspects not explicitly stated in the paper
- Facilitate publishing new papers (others can use the data)
How to Publish a new FL Paper?
How to Publish a new FL Paper?
How to Publish a new FL Paper?

- Datasets
  - Few available
- New FLT
  - Hard to generate
How to Publish a new FL Paper?

Datasets

Results

New FLT
How to Publish a new FL Paper?

Datasets → New FLT

Results

?
How to Publish a new FL Paper?

Datasets

Results

FLT#1  FLT#2  FLT#3  FLT#4  FLT#5  FLT#N

New FLT
How to Publish a new FL Paper?

Datasets

Results

FLT#1  FLT#2  FLT#3  FLT#4  FLT#5  FLT#N

New FLT

No Implementation
How to Publish a new FL Paper?

Datasets

New FLT

Results

FLT#1  FLT#2  FLT#3  FLT#4  FLT#5  FLT#N

No Implementation

Implement from scratch
How to Publish a new FL Paper?

Datasets

New FLT

Results

FLT#1  FLT#2  FLT#3  FLT#4  FLT#5  FLT#N

No Implementation

Implementation

Implement from scratch
How to Publish a new FL Paper?

Datasets

New FLT

Results

FLT#1   FLT#2   FLT#3   FLT#4   FLT#5   FLT#N

No Implementation          Implementation

Different metrics
Different inputs, etc.

Implement from scratch
How to Publish a new FL Paper?

Datasets

New FLT

Results

FLT#1  FLT#2  FLT#3  FLT#4  FLT#5  FLT#N

No Implementation

Implementation

Different metrics
Different inputs, etc.

Implement from scratch

Modify the tool
How to Publish a new FL Paper?

Datasets

New FLT

Results

FLT#1
FLT#2
FLT#3
FLT#4
FLT#5
FLT#N

No Implementation
Implementation
Implementation

Different metrics
Different inputs, etc.

Implement from scratch

Modify the tool
How to Publish a new FL Paper?

Datasets

Results

FLT#1
FLT#2
FLT#3
FLT#4
FLT#5
FLT#N

No Implementation
Implementation
Implementation

Different metrics
Different inputs, etc.
Details missing

Implement from scratch
Modify the tool

New FLT
How to Publish a new FL Paper?

Datasets

New FLT

Results

No Implementation

Implementation

Implementation

Different metrics
Different inputs, etc.

Details missing

Implement from scratch

Modify the tool

“guesstimate”?
How to Publish a new FL Paper?

Tedious and error prone

No Implementation

Implementation

Different metrics
Different inputs, etc.

Details missing

Implement from scratch

Modify the tool

“guesstimate”?
How to Publish a new FL Paper?

Tedious and error prone

Is there a better way?

Implement from scratch

Modify the tool

“guesstimate”?

Different metrics
Different inputs, etc.

Details missing
How to Publish a new FL Paper?

Tedious and error prone

Is there a better way?

TraceLab
TraceLab

• Framework for:
  - Creating
  - Conducting experiments
  - Sharing

• Designed to support traceability link recovery

• Developed by:

  [DePaul University Logo]
  [Kent State University Logo]
  [William & Mary Logo]
TraceLab
TraceLab

Experiment
TraceLab

Components (C#, Java)

Experiment
Components (C#, Java)

TraceLab

Supports scripting languages: R, Python

Experiment
How we Envision Publishing FL Papers with TraceLab?
How we Envision Publishing FL Papers with TraceLab?
How we Envision Publishing FL Papers with TraceLab?
Running the Experiment

Status: Experiment done!
Running the Experiment
How to Create a Composite Component?

Diagram:

- Start
  - Corpus Importer
    - Corpus Remove Non-Literals
    - Corpus Split Identifiers
    - Corpus Stopwords
    - Corpus Stemmer
    - TFIDF Corpus
      - Compute VSM Similarities
        - Effectiveness All Ranks - VSM
        - Effectiveness Best Ranks - VSM
      - End
  - Gold Set Importer
  - Queries Importer
    - Queries Remove Non-Literals
    - Queries Split Identifiers
    - Queries Stopwords
    - Queries Stemmer
    - Effectiveness All Ranks - VSM
    - Effectiveness Best Ranks - VSM
    - End
How to Create a Composite Component?

Define component info:
- Component name: VSM FLT Experiment
- Author: Evan Moritz
- Version:
- Description: Performs the Vector Space Model Feature Location Technique.

Import
- GoldSet: TraceLabSDK.Types.TLSimilarityMatrix
- DocumentMap: System.Collections.Generic.List`1[System.String,mscorlib, Version=4.0.0.0, Culture=neutral, PublicKeyToken=b77a5c561934e089]
- Corpus: TraceLabSDK.Types.TLArtifactsCollection
- Queries: TraceLabSDK.Types.TLArtifactsCollection
- Stopwords: TraceLabSDK.Types.TLStopwords

Output
- Corpus: TraceLabSDK.Types.TLArtifactsCollection
- Queries: TraceLabSDK.Types.TLArtifactsCollection
- listOfArtifacts: TraceLabSDK.Types.TLArtifactsCollection
- VSMSimilarities: TraceLabSDK.Types.TLSimilarityMatrix
- CorpusIndex: TraceLabSDK.Types.TLDictionaryIndex

Config Parameter | Alias | Type
--- | --- | ---
Gold Set Importer Directory | Gold Set Importer Directory | System.String
Corpus Importer Identifiers | Corpus Importer Identifiers | TraceLabSDK.Component.Cc
Corpus Importer Documents | Corpus Importer Documents | TraceLabSDK.Component.Cc
Queries Importer QueriesDir | Queries Importer QueriesDir | System.String
Queries Importer FeaturesIC | Queries Importer FeaturesIC | TraceLabSDK.Component.Cc
Queries Importer BugsIDFile | Queries Importer BugsID | TraceLabSDK.Component.Cc
How to Create a Composite Component?

Export experiment as component:

Define component info:
- Component name: VSM FLT Experiment
- Author: Evan Moritz
- Version:
- Description: Performs the Vector Space Model Feature Location Technique.

Import | Type | Include
--- | --- | ---
GoldSet | TraceLabSDK.Types.TLSimilarityMatrix | ✓
Corpus | TraceLabSDK.Types.TLArtifactsCollection | ✓
Queries | TraceLabSDK.Types.TLArtifactsCollection | ✓
Stopwords | TraceLabSDK.Types.TLStopwords | ✓

Output | Type | Include
--- | --- | ---
Corpus | TraceLabSDK.Types.TLArtifactsCollection | ✓
Queries | TraceLabSDK.Types.TLArtifactsCollection | ✓
listOfArtifacts | TraceLabSDK.Types.TLArtifactsCollection | ✓
VSMSimilarities | TraceLabSDK.Types.TLSimilarityMatrix | ✓
CorpusIndex | TraceLabSDK.Types.TLDictionaryIndex | ✓

Config Parameter | Alias | Type
--- | --- | ---
Gold Set Importer Directory | Gold Set Importer Directory | System.String
Corpus Importer Identifiers | Corpus Importer Identifiers | TraceLabSDK.Component.ColumnName
Corpus Importer Documents | Corpus Importer Documents | TraceLabSDK.Component.ColumnName
Queries Importer QueriesDir | Queries Importer QueriesDir | System.String
Queries Importer FeaturesIC | Queries Importer FeaturesIC | TraceLabSDK.Component.ColumnName
Queries Importer BugsIDFile | Queries Importer BugsID | TraceLabSDK.Component.ColumnName
How to Create a Composite Component?

[Diagram of the process]

Export experiment as component:

Define component info:
Component name: VSM FLT Experiment
Author: Evan Moritz
Version:
Description: Performs the Vector Space Model Feature Location Technique.

Import | Type | Include
--- | --- | ---
GoldSet | TraceLabSDK.Types.TLSimilarityMatrix | ✓
Corpus | TraceLabSDK.Types.TLArtifactsCollection | ✓
Queries | TraceLabSDK.Types.TLArtifactsCollection | ✓
Stopwords | TraceLabSDK.Types.TLSimilarity | ✓

Output | Type | Include
--- | --- | ---
Corpus | TraceLabSDK.Types.TLArtifactsCollection | ✓
Queries | TraceLabSDK.Types.TLArtifactsCollection | ✓
listOfArtifacts | TraceLabSDK.Types.TLArtifactsCollection | ✓
VSMSimilarities | TraceLabSDK.Types.TLSimilarityMatrix | ✓
CorpusIndex | TraceLabSDK.Types.TLDictionary | ✓

Config Parameter | Alias | Type
--- | --- | ---
Gold Set Importer Directory | Gold Set Importer Directory | System.String
Corpus Importer Identifier | Corpus Importer Identifier | TraceLabSDK.Component.Co
Corpus Importer Documents | Corpus Importer Documents | TraceLabSDK.Component.Co
Queries Importer QueriesDir | Queries Importer QueriesDir | System.String
Queries Importer FeaturesIC | Queries Importer FeaturesIC | TraceLabSDK.Component.Co
Queries Importer BugsIDFile | Queries Importer BugsIDFile | TraceLabSDK.Component.Co
Comparing the Results
Comparing the Results

Start

- Import Dataset #1
- Import Dataset #m
Comparing the Results

Start

Import Dataset #1

Import Dataset #m

New FLT

FLT #1

FLT #2

FLT #N
Comparing the Results

Start

Import Dataset #1

Import Dataset #m

New FLT

FLT #1

FLT #2

FLT #N

Compute Results Metric #1

Compute Results Metric #2
Comparing the Results

Start

Import Dataset #1

Import Dataset #m

New FLT

FLT #1

FLT #2

FLT #N

Compute Results Metric #1

Compute Results Metric #2

Box Plots

Descriptive Statistics

Statistical Tests

End
TraceLab Vision for FL

• Platform for creating, reproducing and sharing experiments

• In the future, submit a paper with a link to a TraceLab experiment:
  - Data
  - Code
  - Configuration parameters
  - Etc.
Status of TraceLab

• Beta version available by request
• TraceLab will be released in July 2012

• Features under development:
  - Packaging feature that bundles up:
    • Data
    • Code
    • Settings
  - Performance and GUI enhancements
Our Contribution

http://www.cs.wm.edu/sem eru/data/TraceLab-feature-location/

• Datasets for 5 systems:
  - Features to methods mapping
  - Execution traces
  - Queries
  - Corpora

• Reusable components and sample experiments for FL

• Evaluation metrics
Conclusions and Future Work

• TraceLab & Feature Location = Reproducible experiments

• Research community welcomed to share:
  - More datasets
  - More components and experiments
Conclusions and Future Work

• Adapt TraceLab for other SE areas:
  - Impact analysis
  - Detection of duplicate bug reports
  - Developer recommendation

• At ICSE’12 TraceLab got a lot of interest from large software groups

• Hopefully adaption by ICPC community
Thank you! Questions?

SEMERU @ William and Mary

http://www.cs.wm.edu/semeru/
Generating the jEdit Datasets

---

r13898 | shlomy | 2008-10-17 00:45:12 -0400 (Fri, 17 Oct 2008) | 2 lines
Changed paths:
M /jEdit/trunk/doc/CHANGES.txt
M /jEdit/trunk/org/gjt/sp/jedit/search/HyperSearchRequest.java

_fixed bug #2173172: search within selection of regexp '`.` does not match the last line. This was caused by an exception - if the last selection line is empty, buffer.getSegment() was called with a negative length of -1. Also removed an unused variable._
---

SVN Commits between v4.2-v4.3
Generating the jEdit Datasets

**Title**

**Description**

**Query**

---

 SVN commit message

```sh
r13898 | shlomy | 2008-10-17 00:45:12 -0400 (Fri, 17 Oct 2008) | 2 lines
Changed paths:
M /jEdit/trunk/doc/CHANGES.txt
M /jEdit/trunk/org/gjt/sp/jedit/search/HyperSearchRequest.java
```

Fixed bug #2173712: search within selection of regexp '{' does not match the last line. This was caused by an exception - if the last selection line is empty, buffer.getSegment() was called with a negative length of -1. Also removed an unused variable.
Generating the jEdit Datasets

r13898 | shlomy | 2008-10-17 00:45:12 -0400 (Fri, 17 Oct 2008) | 2 lines

Changed paths:

M /jEdit/trunk/doc/CHANGES.txt
M /jEdit/trunk/org/gjt/sp/jedit/search/HyperSearchRequest.java

Fixed bug #2173172: search within selection of regexp `\.' does not match the last line. This was caused by an exception - if the last selection line is empty, buffer.getSegment() was called with a negative length of -1. Also removed an unused variable.

---

Modified methods (gold set)

59
Configuring Components
Comparing the Results (Alternative)