Revisiting the Experimental Design Choices for Approaches for the Automated Retrieval of Duplicate Issue Reports

Ph.D. Thesis Defense

Ph.D. : Mohamed Sami Rakha
Supervisor: Prof. Ahmed E. Hassan
A problem has been detected and Windows has been shut down to prevent damage to your computer.

PAGE_FAULT_IN_NONPAGED_AREA

If this is the first time you’ve seen this Stop error screen, restart your computer. If this screen appears again, follow these steps:

Check to make sure any new hardware or software is properly installed. If this is a new installation, ask your hardware or software manufacturer for any Windows updates you might need.

If problems continue, disable or remove any newly installed hardware or software. Disable BIOS memory options such as caching or shadowing. If you need to use Safe Mode to remove or disable components, restart your computer, press F8 to select Advanced Startup options, and then select Safe Mode.

Technical information:

*** STOP: 0x00000050 (0x8872A990, 0x0000001)
Software issues are common

PAGE_FAULT_IN_NONPAGED_AREA

If this is the case, restart your computer.

Check to make sure you have the latest Windows updates.
If this is a new application, check for any Windows updates.

If problems continue, stop the application or software. Did you install something before the crash?
If you need to access an application on your computer, select Safe Mode.

Technical information:

*** STOP: 0x00000050 (0x872A990, 0x0000001)***
There are many issue tracking systems

Bugzilla

JIRA

FogBugz

assembla
Issue tracking systems receive a large number of issue reports.
Different persons may report the same issue
Different persons may report the same issue
The identification of duplicate issue reports is an annoying and timewasting task

20-30% Duplicates!

Bettenburg et al. Duplicate bug reports considered harmful… really? [ICSM 2008:]
Information retrieval is leveraged by prior research.
Developer choose the accurate candidate for the newly reported duplicate... If there is any
Several automated approaches for duplicates retrieval are proposed by prior research:

- 2007: Runeson et al.
- 2008: Wang et al.
- 2010: Sun et al.
- 2010: Sureka et al.
- 2011: Sun et al.
- 2012: Nguyen et al.
- 2012: Zhou et al.
- 2012: Nguyen et al.
- 2015: Aggarwal et al.
- 2016: Jie et al.
- 2016: Hindle et al.
- 2016: Jie et al.
Prior research focuses on improving the performance.

The goal of newly proposed approach is to have higher performance than prior ones.

Prior Research
- 2016: Jie et al.
- 2016: Hindle et al.
- 2015: Aggarwal et al.
- 2012: Nguyen et al.
- 2012: Zhou et al.
- 2011: Sun et al.
- 2010: Sureka et al.
- 2010: Sun et al.
- 2008: Wang et al.
- 2007: Runeson et al.
In my thesis, I revisit the experimental design choices from four perspectives. The goal of newly proposed approach is to have better performance than prior ones. Revisiting the experimental design choices can help better understand the benefits and limitations of prior approaches evaluation.
Prior research treats all duplicate reports equally.

Duplicate reports differ based on the effort needed for their manual identification.
Real Example:
Duplicate (B) needs more much effort to be identified
Real Example:
Duplicate (B) needs more much effort to be identified

15 minutes

reported

Identified
Real Example:
Duplicate (B) needs much more effort to be identified.
A duplicate can be **hard** or **easy** based on the effort needed for identification.
Evaluation Process

Data Filtration

Data Changes

Needed Effort

Within 1 Day

Without Discussions

With one Person

Studied Systems:

- EASY ≈ 50%
- HARD ≈ 50%
Evaluation Process

Data Filtration

Data Changes

Random Forest Model

“Binary Classification”

AUCs

0.68 – 0.8

Needed Effort

Fast Identification

EASY ≈ 50%

Slow Identification

HARD ≈ 50%
Random Forest Model “Binary Classification”

AUCs 0.68 – 0.8

This model is used to explain the most important factors on the effort needed for identification.

Factors such as **Description Similarity** is one of the top important ones.
Thesis Contribution #1:
Show the importance of considering the needed effort for retrieving duplicate reports in the performance measurement of automated duplicate retrieval approaches

[Published in EMSE 2015]
Many prior research evaluate their approaches on **short testing periods**

Not considered in the evaluation

Testing period (1 Year)
Many prior research evaluate their approaches on **short testing periods**

- Not considered in the evaluation
- Testing period (1 Year)
Duplicate (B) match is outside the testing period.
Duplicate (B) match is outside the testing period

Not considered in the evaluation

Testing period (1 Year)
I refer to this evaluation by "Classical Evaluation"
Instead, I propose “Realistic Evaluation”
Instead, I propose “Realistic Evaluation”
I compare the two evaluations on frequently used testing data in prior research.

**Problem #1**
Number of considered duplicates

**Problem #2**
Performance overestimation

- Mozilla 2010
- Eclipse 2008
- OpenOffice 2008 - 2010
Problem #1
Number of considered duplicates

Classical evaluation is missing 23.5% of duplicates

Classical
Realistic Evaluation

7,551 Duplicates
10,043 Duplicates

Testing period
Recall rate is overestimated by 39% for REP approach.

- Classical Evaluation
  - Recall: 0.43
  - Recall: 0.58

- Realistic Evaluation
  - Recall: 0.43
  - Recall: 0.58

Mozilla 2010

Problem #2 Performance overestimation

Time

Testing period
Classical Evaluation

Realistic Evaluation

Evaluation Process

Needed Effort

Data Filtration

Data Changes

Classical Evaluation

Realistic Evaluation

Problem #1
Number of considered duplicates

Problem #2
Performance overestimation

Mozilla 2010
Eclipse 2008
OpenOffice 2008 - 2010
Is the difference in performance between the two evaluations generalize to other testing periods within the tested ITs?
Randomly select 100 chunks as “Testing Periods”

Studying the statistical difference between the classical and realistic evaluations

Chunk = 1 Year

Classical Evaluation

Realistic Evaluation
Classical Evaluation

Realistic Evaluation

Chunk 1 (Dec2004 – Dec2005)


Chunk 100 (Sep2007 – Sep2008)

Evaluation Process

Data Changes

Data Filtration

Needed Effort

Results 1

Results 2

Automated Approach

Results 100
Classical evaluation significantly overestimates the performance of the automated approaches (p-value < 0.001)
Thesis Contribution #2: Propose a realistic evaluation for the automated retrieval of duplicate reports, and analyze prior findings using such realistic evaluation

[Published in TSE 2017]
Do we need all these issue reports to search in?

Filtering out old issue reports should lead to reduced search time.
The possible solution is to filter out the old issue reports using a time window based on the report date. No duplicates are ignored.
The possible solution is to filter out the old issue reports using a time window based on the report date.
The possible solution is to filter out the old issue reports using a time window based on the report date.
Realistic Evaluation
It is **not possible** to limit the issue reports that are searched **without decreasing** the **performance** of automated approaches.
Data Changes

Needed Effort

Evaluation Process

Data Filtration

Time

Testing period

reported
(A) is resolved 1 month ago.
(B) is resolved 2 years ago

(A) is resolved 1 month ago

Testing period
A new duplicate is more likely to be a duplicate of (A) as (B) is resolved 2 years ago and (A) is resolved 1 month ago.

Testing period
I propose **filtration approach** based on **time threshold** from resolution date for each resolution **value** and **rank**.

The thresholds are optimized using **Genetic Algorithm (NSGA-II)**.
The proposed filtration approach improves the performance by up to 60%
Thesis Contribution #3:
Propose a genetic algorithm based approach for filtering out the old issue reports to improve the performance of duplicate retrieval approaches.

[Published in TSE 2017]
Just-in-time (JIT) duplicate retrieval feature is introduced
### Summary

**Needed Effort**

**Evaluation Process**

**Data Filtration**

**Data Changes**

**NEW FEATURE!**

<table>
<thead>
<tr>
<th>Bug ID</th>
<th>Summary</th>
<th>Status</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>37277</td>
<td>Title Area Dialog requires navigation to the error message</td>
<td>CLOSED FIXED</td>
<td>Add Me to the CC List</td>
</tr>
<tr>
<td>42176</td>
<td>[Navigator] Invalid thread access when closing a project</td>
<td>RESOLVED FIXED</td>
<td>Add Me to the CC List</td>
</tr>
<tr>
<td>73238</td>
<td>[navigation] Create a ctrl+click navigation extension point in editor</td>
<td>ASSIGNED</td>
<td>Add Me to the CC List</td>
</tr>
<tr>
<td>75271</td>
<td>Tab and some of it's composite Accesskey error in navigation view.</td>
<td>RESOLVED FIXED</td>
<td>Add Me to the CC List</td>
</tr>
<tr>
<td>165527</td>
<td>[Navigation] Backwards navigation across elements doesn't select last diff</td>
<td>ASSIGNED</td>
<td>Add Me to the CC List</td>
</tr>
<tr>
<td>321833</td>
<td>[EditorMgmt] [navigation] Navigator Stack does not work with multiple editors on same file</td>
<td>NEW</td>
<td>Add Me to the CC List</td>
</tr>
<tr>
<td>404535</td>
<td>[EditorMgmt] Enabling navigation history for inter tab navigation in multi page editors</td>
<td>NEW</td>
<td>Add Me to the CC List</td>
</tr>
</tbody>
</table>
Significantly less duplicates but more identification effort is needed.
<table>
<thead>
<tr>
<th>Needed Effort</th>
<th>Evaluation Process</th>
<th>Data Filtration</th>
<th>Data Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

How the **JIT** duplicate retrieval feature **impact** the **performance evaluation** of the automated approaches?
The performance of both approaches is significantly lower for after-JIT duplicates.
The relative improvement of the performance of REP over BM25F is significantly larger after the activation of the JIT.
Thesis Contribution #4:
Highlight the **impact** of the “**JIT**” feature on the **performance evaluation** of duplicate retrieval approaches

[Submitted to EMSE] – under review
Summary

Prior research focuses on improving the performance

The goal of newly proposed approach is to have better performance than prior ones

Example:
A > B
Recall Rate = 0.6
Recall Rate = 0.5

In my thesis, I revisit the experimental design choices from four perspectives

Revisiting the experimental design choices can help better understand the **benefits** and **limitations** of prior approaches evaluation.

**Thesis Contribution #1:**
Show the importance of considering the needed effort for retrieving duplicate reports in the performance measurement of automated duplicate retrieval approaches.

[EMSE 2015]

**Thesis Contribution #4:**
Highlight the impact of the “JIT” feature on the performance evaluation of duplicate retrieval approaches.

[EMSE 2017] – under review
Prior research focuses on improving the performance.

The goal of newly proposed approach is to have higher performance than prior ones.

Prior Research:

- 2016: Jie et al.
- 2016: Hindle et al.
- 2015: Aggarwal et al.
- 2012: Nguyen et al.
- 2012: Zhou et al.
- 2011: Sun et al.
- 2010: Sureka et al.
- 2010: Sun et al.
- 2008: Wang et al.
- 2007: Runeson et al.
In my thesis, I revisit the experimental design choices from four perspectives:

**Revisiting** the experimental design choices can help better understand the **benefits** and **limitations** of prior approaches **evaluation**.

**Needed Effort**

**Evaluation Process**

**Data Filtration**

**Data Changes**

Recall Rate = 0.6
Recall Rate = 0.5

Example:

2007: Runeson et al.
2008: Wang et al.
2010: Sun et al.
2010: Sureka et al.
2011: Sun et al.
2012: Zhou et al.
2012: Nguyen et al.
2015: Aggarwal et al.
2016: Hindle et al.
2016: Jie et al.
Thesis Contribution #1:
Show the importance of considering the needed effort for retrieving duplicate reports in the performance measurement of automated duplicate retrieval approaches

[Published in EMSE 2015]
Thesis Contribution #2:
Propose a realistic evaluation for the automated retrieval of duplicate reports, and analyze prior findings using such realistic evaluation

[Published in TSE 2017]
Thesis Contribution #3:
Propose a genetic algorithm based approach for filtering out the old issue reports to improve the performance of duplicate retrieval approaches

[Published in TSE 2017]
Thesis Contribution #4:
Highlight the **impact** of the “**JIT**” feature on the **performance evaluation** of duplicate retrieval approaches

[EMSE 2017] – under review
Summary

Prior research focuses on improving the performance

The goal of newly proposed approach is to have better performance than prior ones

Example: A > B
Recall Rate = 0.6
Recall Rate = 0.5

Prior Research
2016: Jie et al.
2016: Hindle al.
2015: Aggarwal al.
2012: Nguyen al.
2012: Zhou al.
2011: Sun et al.
2010: Sureka et al.
2010: Sun et al.
2008: Wang et al.
2007: Runeson et al.

In my thesis, I revisit the experimental design choices from four perspectives

Revisiting the experimental design choices can help better understand the benefits and limitations of prior approaches evaluation.

Thesis Contribution #1:
Show the importance of considering the needed effort for retrieving duplicate reports in the performance measurement of automated duplicate retrieval approaches.

[EMSE 2015]

Thesis Contribution #4:
Highlight the impact of the “JIT” feature on the performance evaluation of duplicate retrieval approaches.

[EMSE 2017] – under review