Duplicate Bugs Detection and Retrieval in Software Projects : Commentary and Recommendations

Neela Krishna Teja Tadikonda

Masters Student

Department of Computer Science

North Carolina State University  
Raleigh, North Carolina  
Mobile - +1-9195375046

ntadiko@ncsu.edu

**ABSTRACT**

In this paper, we describe the formatting guidelines for ACM SIG Proceedings.

**CCS Concepts**

• **Information systems➝Database management system engines**   • **Computing methodologies➝Massively parallel and high-performance simulations.**This is just an example, please use the correct category and subject descriptors for your submission*.* The ACM Computing Classification Scheme:

<http://www.acm.org/about/class/class/2012>. Please read the [HOW TO CLASSIFY WORKS USING ACM'S COMPUTING CLASSIFICATION SYSTEM](http://www.acm.org/publications/article-templates/CCS-HOWTO-v6-12Jan2015.docx) for instructions on how to classify your document using the 2012 ACM Computing Classification System and insert the index terms into your Microsoft Word source file.

**Keywords**

Keywords are your own designated keywords separated by semicolons (“;”).

# INTRODUCTION

The sooner we find the software defects the sooner we can fix them. Find bugs sooner and fixing them is very essential to the software quality assurance. It can reduce cost that is induced if bugs were found later than earlier. It is often the case that a same fault in the software product is reported in multiple perspectives by different people. Usually their vantage points are different but the issue they are reporting could be same. Not every project is funded well in man power or tools, developers fixing these bugs need to invest their time, which is a valuable resource, in the software engineering discipline. Identifying duplicate bugs is therefore essential for the developer so that they plan the efforts and efficiently fix the bugs. Over the last decade There is considerable amount of effort spent by the researchers over finding duplicate bugs.

# Related Work

Nguyen [1] used unsupervised Machine Learning Techniques to detect the duplicate bugs , such as Topic Modelling and

i ii 1 Techincal issue: The actual root cause of a bug. Various bugs can share a single Technical issue. In other words two bug reports can share similar topics. 2 Technical topic: Perspective based reports with different description of the root cause, suggestions, tips for the fixes e.t.c with characteristics of latent, semantic features. 3 Latent: Details that are not obvious but inherent in the context and understood by comprehension. 4 Ensemble Averaging: It is Linear weighted averaging of candidates. In this paper DBTM and BM25F a combined using Ensemble Averaging. iii 1 Motivational Statements Two bug reports addressing same issue, but describe it from two different perspectives. This adds common technical topics between them among other technical topics reported. 2 Related Work Provided previous work on this problem - some ML techniques like Information Retrieval Model, Vector Space Model using simple Natural language Processing, binary classification ( applying linear regression ), Support Vector machine, extended BM25F known as REP. 3 Informative Visualizations: Figures showing the variables derived in Topic Model with dependencies between them. Some figures showing the accuracy comparisons across several models, over various sources of bug reports. 4 Baseline Results: Results comparing the accuracy of the DBTM against previous models like REP, T-Model, BM25F, and RTM+BM25F over various bug report sources like OpenOffice, Mozilla and Eclipse. Where DBTM displays higher accuracy than other models at all places. 5 Sampling Method: Gibbs sampling method. It is used for training the DBTM with identified historical bug report data. 6 Patterns: The DBTM sensitivity is analyzed by varying number of topics from 20 to 400 in steps of 10 and measured the top 10 accuracy. 7 Negative Results: Using less number of topic gives less accuracy as the comparison results in more bugs going into one duplicate group. iv 1 To further improve the time efficiency, bug report groups can be strategically chosen for comparison with bug report groups. For example, we can use the topic proportions for a new bug report, against the bug report groups topics to strategically pick bug report groups tp do the comparison, instead of going bruteforce with each new bug on the group. 2 The paper can be improved by providing any further enhancements that could have been made by them to improve the efficiency, atleast a line or 2 on the open ended questions and future work. 3 A quick overview on the preproessing of the text document into words v two A. T. Nguyen, T. T. Nguyen, J. Al-Kofahi, H. V. Nguyen, and T. N. Nguyen. A Topic-based Approach for Narrowing the Search Space of Buggy Files from a Bug Report. In ASE11, pp. 263-272. IEEE CS, 2011. – This paper presents an LDA approach to finding the bug reports corresponding to a source file. three C. Sun, D. Lo, S.-C. Khoo, and J. Jiang. Towards more accurate retrieval of duplicate bug reports. In ASE11, pages 253262. IEEE CS, 2011. – This paper develops an extended BM25F technique to find the duplicate bugs and presents its results. four C. Sun, D. Lo, X. Wang, J. Jiang, and S.-C. Khoo. A discriminative model approach for accurate duplicate bug report retrieval. In ICSE10. ACM, 2010. – This paper also talks about the duplicate bug identification using Support Vector machine for making a discriminative model.

All material on each page should fit within a rectangle of 18 × 23.5 cm (7" × 9.25"), centered on the page, beginning 1.9 cm (0.75") from the top of the page and ending with 2.54 cm (1") from the bottom. The right and left margins should be 1.9 cm (.75"). The text should be in two 8.45 cm (3.33") columns with a .83 cm (.33") gutter.

# TYPESET TEXT

## Normal or Body Text

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

*Conference’10*, Month 1–2, 2010, City, State, Country.

Copyright 2010 ACM 1-58113-000-0/00/0010 …$15.00.

DOI: http://dx.doi.org/10.1145/12345.67890

Please use a 9-point Times Roman font, or other Roman font with serifs, as close as possible in appearance to Times Roman in which these guidelines have been set. The goal is to have a 9-point text, as you see here. Please use sans-serif or non-proportional fonts only for special purposes, such as distinguishing source code text. If Times Roman is not available, try the font named Computer Modern Roman. On a Macintosh, use the font named Times. Right margins should be justified, not ragged.

## Title and Authors

The title (Helvetica 18-point bold), authors' names (Helvetica 12-point) and affiliations (Helvetica 10-point) run across the full width of the page – one column wide. We also recommend phone number (Helvetica 10-point) and e-mail address (Helvetica 12-point). See the top of this page for three addresses. If only one address is needed, center all address text. For two addresses, use two centered tabs, and so on. For more than three authors, you may have to improvise.[[1]](#footnote-1)

## First Page Copyright Notice

Please leave 3.81 cm (1.5") of blank text box at the bottom of the left column of the first page for the copyright notice.

## Subsequent Pages

For pages other than the first page, start at the top of the page, and continue in double-column format. The two columns on the last page should be as close to equal length as possible.

Table . Table captions should be placed above the table

|  |  |  |  |
| --- | --- | --- | --- |
| **Graphics** | **Top** | **In-between** | **Bottom** |
| Tables | End | Last | First |
| Figures | Good | Similar | Very well |

## References and Citations

Footnotes should be Times New Roman 9-point, and justified to the full width of the column.

Use the “ACM Reference format” for references – that is, a numbered list at the end of the article, ordered alphabetically and formatted accordingly. See examples of some typical reference types, in the new “ACM Reference format”, at the end of this document. Within this template, use the style named *references* for the text. Acceptable abbreviations, for journal names, can be found here: <http://library.caltech.edu/reference/abbreviations/>. Word may try to automatically ‘underline’ hotlinks in your references, the correct style is NO underlining.

The references are also in 9 pt., but that section (see Section 7) is ragged right. References should be published materials accessible to the public. Internal technical reports may be cited only if they are easily accessible (i.e. you can give the address to obtain the report within your citation) and may be obtained by any reader. Proprietary information may not be cited. Private communications should be acknowledged, not referenced (e.g., “[Robertson, personal communication]”).

## Page Numbering, Headers and Footers

Do not include headers, footers or page numbers in your submission. These will be added when the publications are assembled.

# FIGURES/CAPTIONS

Place Tables/Figures/Images in text as close to the reference as possible (see Figure 1). It may extend across both columns to a maximum width of 17.78 cm (7”).

Captions should be Times New Roman 9-point bold. They should be numbered (e.g., “Table 1” or “Figure 2”), please note that the word for Table and Figure are spelled out. Figure’s captions should be centered beneath the image or picture, and Table captions should be centered above the table body.

# SECTIONS

The heading of a section should be in Times New Roman 12-point bold in all-capitals flush left with an additional 6-points of white space above the section head. Sections and subsequent sub- sections should be numbered and flush left. For a section head and a subsection head together (such as Section 3 and subsection 3.1), use no additional space above the subsection head.



Figure 1. Insert caption to place caption below figure.

.

## Subsections

The heading of subsections should be in Times New Roman 12-point bold with only the initial letters capitalized. (Note: For subsections and subsubsections, a word like *the* or *a* is not capitalized unless it is the first word of the header.)

### Subsubsections

The heading for subsubsections should be in Times New Roman 11-point italic with initial letters capitalized and 6-points of white space above the subsubsection head.

#### Subsubsections

The heading for subsubsections should be in Times New Roman 11-point italic with initial letters capitalized.

#### Subsubsections

The heading for subsubsections should be in Times New Roman 11-point italic with initial letters capitalized.

# ACKNOWLEDGMENTS

Our thanks to ACM SIGCHI for allowing us to modify templates they had developed.

# REFERENCES

1. Nguyen, A. T., Nguyen, T. T. T. N., Lo, D., and Sun, C. (2012). Duplicate bug report detection with a combination of information retrieval and topic modeling. Proceedings of the 27th IEEE/ACM International Conference on Automated Software Engineering - ASE 2012, 70. Retrieved from [http://dl.acm.org/citation.cfm?doid= 2351676.2351687](http://dl.acm.org/citation.cfm?doid=%202351676.2351687)

Columns on Last Page Should Be Made As Close As Possible to Equal Length

1. If necessary, you may place some address information in a footnote, or in a named section at the end of your paper. [↑](#footnote-ref-1)