Duplicate Bug Report Rate in Oss Projects: A Comparative Analysis

Swati Sen[1], AnitaGanpati[2], AmanKumar Sharma[3]

[1]*Research Scholar, Department of Computer science, Himachal Pradesh University, Shimla,*

[2]*Assistant Professor,Department of Computer science,Himachal Pradesh University, Shimla,*

[3]*Associate Professor, Department of Computer Science, Himachal Pradesh University, Shimla,*

,

Abstract- *The use of bug tracking system in Open Source Software (OSS) to organize maintenance activity of software is wide spread. Bug tracking system is open bug repository that is maintained by open source software organizations to track their bugs so that bug reports from all over the world can be gathered.To improve the reliability of software system developers often allow user to submit a bug by writing bug report in bug tracking system over the internet. The system of tracking bugs in open bug repository is totally distributed and uncontrolled. Different reporters may submit same bug report again and again for the sameproblem. The same report which is submitted by several reporters is referred as duplicate bug report. Identification of these duplicate reports is time consuming and intensifies the already high cost of software maintenance. Duplicate bug put extra overhead on software organizations, as they put negative effect on maintenance of software. So utility of these systems is hindered by excessive number of duplicates. In this study bug repository of open source projects was explored to find out whether there is enough number of duplicate bugs to cause the problem for efficient maintenance of software. To show the extent of duplication problem duplicate bug rate was calculated by dividing the number of duplicate bug with total number of bugs in six open source project's bug tracking system. Project under study were Thunderbird, Mandriva Linux, and Firefox for Android, Eclipse BIRT, Penelope and Kompare. It is evident from the result that duplication problem is found in each studied project. Theduplication rate is not constant among all the studied projects.The average rate is 25.3% which is large enough to cause problem in OSS project's maintenance activity.*

Keywords-*Bug Tracking System, Open Source Software, Reporter, Developers, Bug Report, Duplicate Bug Report, Maintenance, Duplicate Bug Rate.*

1.Introduction

Open SourceSoftware (OSS) has been gaining attention in the last few years. Many corporations, large or small, have shown an interest in this growing open source software market. This shows some strong differences with traditional software. Unlike traditional software OSS is developed by distributed teams through the internet. Developers of OSS contribute from all around the world for activity of development.Developers rarely or never meet face-to-face and coordinate their activity on internet. After development maintenance of software is required to meet the desired effect. Almost all development organization track software bug to guide the maintenance activity of software developers. Bugs are prevalent and widespread in software system.A large part of software development and maintenance is spent on locating and fixing bugs. The use of the bug tracking tool to organize maintenance activity in software is widespread.To improve the reliability of software system developers allow user to provide feedback on bug by writing bug report in bug tracking system like Bugzilla, Jiraand Mantis etc. This process is distributed and uncoordinated. Many reporters could submit same bug which is reporting same problem. The same report which is submitted by several reporters is referred to as duplicate bug report. Handling of bug report is important issue in open source software domain because identification of these duplicate reports is time consuming and intensifies the already high cost of software maintenance.In open source domain where large bug repository exist searching the bug tracking system to determine whether a problem has already been submitted is usually higher than the cost of creating a new bug report[1]. The staff responsible for the

search and analysis of bug report must spend a large amount of time and effort. The impact of duplication can be seen in almost all maintenance activity [7]. To show the extent of this problem we explore the bug tracking system of six projects and find out whether the duplication is enough to cause the problem.

A. Bug Tracking System

To manage the process of writing a bug report, solving a bug and communicating the solution to fix a bug communities have created and adopted different system. Most of the project uses bug tracking system.Bug tracking system gives users an opportunity to report and describe failure incidents.A bug tracking system is an application that lets one to keep track of bugs for software project in database. Reporters of bug submit a bug report into bug tracking system which serves as central repository for monitoring the progress of bug. In bug tracking system assignee of the bug fixes it. The role of bug tracking systems in software development is vital.Bug tracking system allows people anywhere in the world to report and describe the bug whenever they like. Bug tracking system usually consists of database known as bug repository. Anyone can access the repository of bug in bug tracking system of project. Rex Black defined bug tracking system that a bug tracking system is some program or application that allow the project team to report, manage and analyze bug report and bug trend. Functionally most bug tracking system provide a form that allow us reporter to report and manage specific bug, a set of stored report and graph that allow us to analyze, manipulate and output this bug data in various ways[2].

B. Bug Report

In open source software environment, user of open source software often write a "bug report" when they find bug or come across a mistake. Bug in bug tracking systems are tracked through bug reports. The report of the bug should be cleared in precisely described. Such report involves the summary, description, version of software in which bug found, severity of bug, OS, priority of bug etc. Bug reports are detailed natural language descriptions of issues or bug. Almost all open source project is supported by an open bug repository in which anyone could have user name and password and either reports an issue or bug or put comment on existing bug report. Bug tracking system usually consists of database known as bug repository which contains information about bug [3]. Bugs can be found in any software or document and these found bugs are reported by submitting bug reports by reporter. These bug reports can be submitted by anyone who finds them.

Peter Farrell Vinay described bug report is used to document any event that occur during testing process which require investigation and layout the minimum requirement for bug management system [4].

C. Duplicate Bug Report

Duplicate report is a report which is submitted by several reporters for same failure or the same defect. There are many users interacting with a system and reporting its issues. Thus the bug is occasionally reported by more than one reporter resulting in duplicate bug report [5]. Whenever developers detect a duplicate, they resolve the report as DUPLICATE and add reference to the master report. In the XML export, the reference of duplicate bug is saved as duplicate identification number, which is used to identify the duplicate bug report and their master bug report [6].Detection of new bug report as a duplicate bug report is critical task. The impact of duplication can be seen in almost all maintenance and evolution activity. To handle these reports a triage need to manually label these bug report as duplicate. Due to the large number of bug reports present in bug repository, it is challenging for the triage to examine all existing bug reports to detect duplication.Triage manually handles these reports and links them to master bug report. The task of duplicate identification is very time consuming and have negative impact on maintenance and productivity.

The main consequence of this problem is the extra effort necessary to detect and manage these duplicate.

2. Literature Survey

In previous work Y.C Cavalcanti et al. explained [7] that according to recent work, duplicate bug report entries in bug tracking systems impact negatively on software maintenance and evolution productivity. In some cases the increased time spent on report analysis and validation, take over 20 minutes. Therefore, a considerable amount of time is lost mainly with duplicate bug report analysis. The work presented an initial characterizationstudy using data from bug trackers from private and open source projects, in order to understand the possible factors that cause bug report duplication and its impact on software development**.**

J.L Davidson et al. [8] explained that Free/Open Source Software (FOSS) communities often use open bug reporting to allow users to participate by reporting bugs. This practice can lead to more

duplicate reports, as users can be less rigorous about researching existing bug reports. The paper examined how FOSS projects deal with duplicate bug reports. They examined 12 FOSS projects: 4 small, 4 medium and 4 large, where size was determined by number of code contributors. First, they found that contrary to what has been reported from studies of individual large projects like Mozilla and Eclipse, duplicate bug reports are a problem for FOSS projects, especially medium-sized, which struggle with a large number of submissions without the resources of large projects. Second, they found that the focus of a project does not affect the number of duplicate bug reports. Their findings indicate a need for additional scaffolding and training for bug reporters.

J. Lerch and M. Mezini [9] addressed that in bug tracking system multiple bug reports are committed for the same bug, which, if not recognized as duplicates, can result in work done multiple times by the development team. Duplicate recognition is, in turn, tedious, requiring examining large amounts of bug reports.

Chengnian Sun et al. explained [10] that in a bug tracking system, different testers or users may submit multiple reports on the same bugs, referred to as duplicates, which may cost extra maintenance efforts in triaging and fixing bugs. In order to identify such duplicates accurately, they proposed a retrieval function (REP) to measure the similarity between two bug reports.

Mehdi Amoui et al. [11] addressed that duplicate defects put extra overheads on software organizations. The cost and effort of managing duplicate defects are mainly redundant. Due to the use of natural language and various ways to describe a defect, it is usually hard to investigate duplicate defects automatically. This problem is more severe in large software organizations with huge defect repositories and massive number of defect reporters.

Yuan Tian et al. [12] explained that the existence of many duplicate bug reports may cause much unnecessary manual efforts as often a triage would need to manually tag bug reports as being duplicates. Recently, there have been a number of studies that investigate duplicate bug report problem which in effect answer the following question: given a new bug report, retrieve other similar bug reports.

Tomi Prifti et.al [13] explained that Bug Tracking Repositories, such as Bugzilla, are designed to support fault reporting for developers, testers and users of the system. Allowing anyone to contribute finding and reporting faults has an immediate impact on software quality. However, this benefit comes with at least one side-effect. Users often file reports that describe the samefault. This increases the maintainer's triage time, but important information required to fix the fault is likely contributed by different reports.

Anh Tuan Nguyen et.al [5] said that detecting duplicate bug reports helps reduce triaging efforts and save time for developers in fixing the same issues. They proposed text-based Information Retrieval (IR) approach which has been shown to outperform others several approaches in term of both accuracy and time efficiency.

3. Need and Scope of Study

Handling of duplicate bug report is an important issue in open source software domain to maintain the efficiency in development and maintenance of software. Due to the large number of bug reports present in bug repository, it is challenging for the triage to examine all existing bug reports to detect duplication. This task is very time consuming and have negative impact on maintenance and productivity of software. Thus due to duplicate bug entry in bug tracker considerable amount of time is lost in duplicate bug report analysis. Keeping the importance of this issue in mind we did exploratory study of duplicate bug in bug tracking system to calculate the duplicate bug rate in different open source software.

In this paper six different projects were studied for the analysis. All project under study use Bugzilla as bug tracker to track their bugs for maintenance activity.

4. Objective of the Study

The objective behind this research is to bring attention towards the increasing rate of duplicate bugs report in bug tracking system of open source software that effects maintenance efficiency in terms of time and effort consumed in bug identification and reporting. This research brings statistical facts about duplicate bug reports. The broad objective of our study is to explore the duplicate bugs in bug repository of bug tracking system and to calculate the duplicate bug rate in different open source project.

5. Analysis

The research was conducted over data collected from bug tracking system shared by developers and reporters for keeping record of software bugs reported during multiple cycles of software development. This work is done to find out that whether the projects have duplicate bug reports in their bug repository and whether the duplicates are large enough to cause problem as duplicate bugs can

increase the maintenance effort. This can be measured by analyzing the bug repository and status of bugs in bug repository of projects under study. If the status of bug is defined or given as duplicate, then the bug is counted as a duplicate bug. A total of six open source projects with their separate bug repositories were considered. Only those bugs were considered for duplication that has been marked as duplicate status. Project under study are Kompare, Eclipse BIRT, Thunderbird, Penelope, Firefox for Android and Mandriva Linux. All projects under study are open source software and use Bugzilla as their bug tracking system.

Table 1: Overview of Projects under Study

Projects	***Bug Tracking System***	***Category Of Project***	***Life Time In Year***
Kompare	*Bugzilla*	*Graphic viewer*	*13*
Eclipse BIRT	*Bugzilla*	*Eclipse based reporting system*	*9*
Thunderbird	*Bugzilla*	*Email application*	*16*
Penelope	*Bugzilla*	*Documentation project*	*7*
Firefox for Android	*Bugzilla*	*Web browser for android*	*2*
Mandriva Linux	*Bugzilla*	*Linux distribution*	*15*

Table 1 gives the overview of project under study which gives type, bug tracking system used by projects, category and lifetime of project in year. In this paper projects were analyzed to collect the data from bug repository of open source projects under study.The total number of bugs and duplicate bugs in bug repository of each project is collected for the lifetime of each project. On the basis of data collected from particular projects bug tracking system the rate of duplicate bugs was calculatedfor each project. Each project under study uses Bugzilla as their bug tracking system to repot bug from all over the world. Rate of duplicate bug in particular project's bug repository can be calculated as:

Rate of duplicate bug = (Number of Duplicate Bugs ÷Total number of bugs) ... *(1)*

Table 2presents total number of bugs, duplicate bugs for all six projects. The collected data provides a duplication rate calculated using formula (1)

Table 2: Bug Tracking System Data

Projects	***Duplicate Bugs***	***Total Number Of Bugs***	***Rate Of Duplicate Bug***
Kompare	*63*	*226*	*28*
Eclipse BIRT	*1896*	*19634*	*9*
Thunderbird	*10438*	*26867*	*39*
Penelope	*144*	*512*	*29*
Firefox for Android	*1942*	*7655*	*26*
Mandriva Linux	*7569*	*35851*	*21*

From Table 2 it is clear that all analyzed project haveproblem of bug duplication at project level. Thunder bird has 10438 duplicate bugs which is highest among all projects. Kompare has 63 duplicate bugs which is lowest among all projects. But rate of duplicate bug is highest in thunderbird and lowest in eclipse BIRT. It shows that duplication depends upon the total number of duplicate bug present in bug repository of particular project under study.

Figure 1 shows the relationshipbetween total number of bugs and duplicate bugs submitted in bug tracking system for particular project.

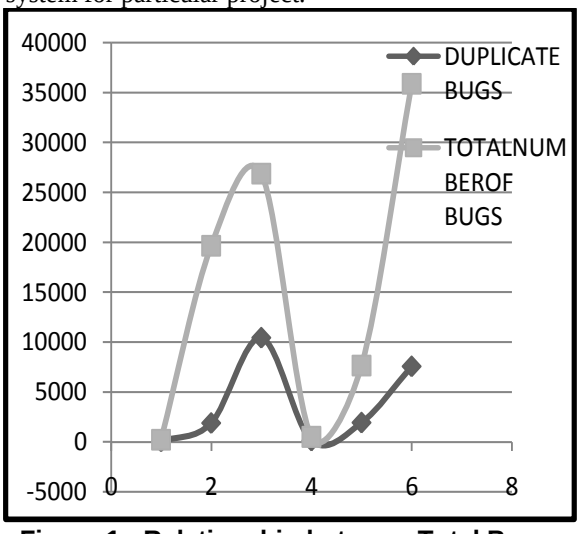

Figure 1: Relationship between Total Bugs and Duplicate Bugs

Figure 1 shows a strong positive relationship between the two counts. This is clear from Figure 1 thatthe increased number of reported bugs will also increase the number of duplicate bugs and decrease in number of total bugs that also decrease the number of duplicate bugs in bug repository. Figure 2 presents a bar chart for rate of duplicate bugs calculated from six project's bug repository.

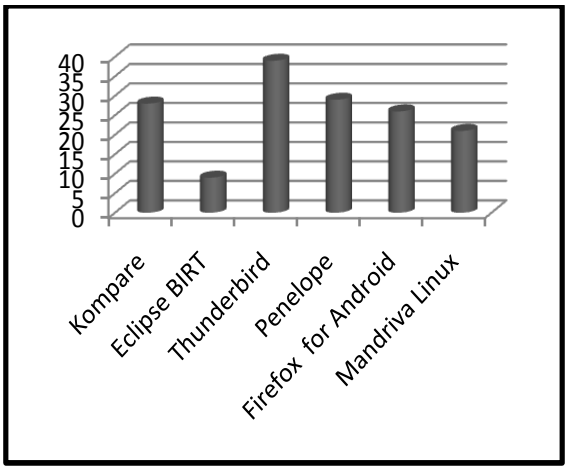

Figure 2: Duplication Rate in Each Project

It is evident from the study that all analyzed project have duplication rate more than 20% with exception in Eclipse BIRT. Thunderbird has 39% rate of duplicate bug which is highest among all project under study. Eclipse BIRT has 9% rate of duplicate bug which is lowest among all studied projects.Rate of duplication is not constant among all projects. The average duplication rate is 25.3% which can be quite significant for considering it as a percentage of overhead for the maintenance schedule in terms of time and cost.

6. Conclusions and Future Work

The study presented extent of duplication problem in six open source software. The extent of duplication is investigated by calculating the rate of duplicate bugs in each six projects. It is evident from the study that each studied projects have duplication problem with highest duplicate rate of 39% in thunderbird and lowest with 9% in eclipse BIRT. Paper also presented relationship between total number of bugs and total number of duplicate bug submitted in bug tracking system of each studied project. There is strong positive relationship between the two counts.It is evident from literature survey that productivity and maintenance being affected by these duplicate bug problem. In future, studies on factors which have impact on duplication and an approach for duplicate bugs in open source project's bug repository can be carried out.

References

[1]Xiaoyin Wang, Lu Zhang, Tao Xie, John Anvik and Jiasu Sun, "An Approach To Detecting Duplicate Bug Reports Using Natural Language And Execution Information", *International Conference on Software Engineering, IEEE,* 2008.

[2]Rex Black, "Managing The Testing Process: Practical Tool And Techniques For Managing Hardware And Software Tool", *John Wiley &sons*, 2003.

[3] Par J. Agerfalk, Cornelia Boldyerf, Jessus M. Gonzolez, Gregory R.Madey and John Noll, "Open Source Software: New Horizons" *springer*, 2010.

[4]] Peter-Farrell-Vinay, "Manage Software Testing", *CRC press*, 2008.

[5] Anh Tuan Nguyen, Tung Thanh Nguyen, T.N. Nguyen and D.Lo, "Duplicate Bug Report Detection With A Combination Of Information Retrieval And Topic Modeling",*International Conference on Automated Software Engineering, Proceedings of the IEEE/ACM*, 2012.

[6]Nicolas Bettenburg, Rahul Premraj, Thomas Zimmermann and Sunghun Kim, "Duplicate Bug Reports Considered Harmful . . . Really?", *IEEE International Conference on Software Maintenance*, 2008.

[7] Y.C Cavalcanti, E. Sade Almeida, Cunha da and D Lucredio, "An Initial Study On The Bug Report Duplication Problem", *Software Maintenance and Reengineering,IEEE,* 2010.

[8] J.L Davison, Nitin Mohan and Carlos Jensen, "Coping with Duplicate Bug Reports In Free/Open Source Software Projects",*Languages and Human-Centric Computing, IEEE,* 2011.

[9] J. Lerch and M. Mezini, "Finding Duplicates OfYour Yet Unwritten Bug Report", *Software Maintenance and Reengineering (CSMR), IEEE*, 2013.

[10] Chengnian Sun, D Lo, Siau-Cheng Khoo and Jing Jiang,"Towards More Accurate Retrieval Of Duplicate Bug Reports*", Automated Software Engineering , IEEE*, 2011.

[11]Mehdi Amoui, Nilam Kaushik, Abraham Al-Dabbagh, Ladan Tahvildari, Shimin Li and Weining Liu, "Search-Based Duplicate Defect Detection: An Industrial Experience" *IEEE*, 2008.

[12] Yuan Tian, Chengnian Sun and David Lo, "Improved Duplicate Bug Report Identification", *Conference on software maintenance and Reengineering, IEEE*, 2012.

[13]Tomi Prifti, Sean Banerjee and Bojan Cukic, "Detecting Bug Duplicate Reports Through Local References",*Proceedings of International Conference on Predictive Models in Software Engineering, ACM*,2011.