Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 3, March 2014, pg.549 – 556

RESEARCH ARTICLE

AN EXPLORATORY STUDY OF DUPLICATE BUG REPORTS IN OSS PROJECTS

Swati Sen[1], Anita Ganpati[2], Aman Kumar Sharma[3]

[1]*Research Scholar, Department of Computer science, Himachal Pradesh University, Shimla, swati.sen12@gmail.com

[2]*Assistant Professor, Department of Computer science, Himachal Pradesh University, Shimla, anitaganpati@gmail.com

[3]*Associate Professor, Department of Computer Science, Himachal Pradesh University, Shimla, sharmaas1@gmail.com

Abstract- Open Source Software (OSS) uses open bug repository during development and maintenance, so that both developer and user can reports bugs that they have found. These systems are generally called as bug tracking system or bug repositories. Bug tracking system is open bug repository that is maintained by open source software organizations to track their bugs. In OSS bug reports from all over the world are gathered which is submitted by geographically distributed team through the use of the internet. Team member of OSS typically works in distributed environment, so the system of tracking bugs in open bug repository is totally distributed and uncontrolled. In OSS different reporters may submit same bug report again and again for the same problem. The same report which is submitted by several reporters is referred to as duplicate bug report. Excessive duplicate bug put extra overhead on software organizations. Utility of software is hindered by these duplicate bugs. In this study bug repository of open source projects was explored to find out the factors that have impact on duplicate bug reports. To find out the factors that have impact on duplicate bug reports bug repository of six open source software project was explored. Factors analyzed for study were numbers of submitters, bug repository size, project size, life span of project and number of developers. Project studied were Thunderbird, Mandriva Linux, and Firefox for Android, Eclipse BIRT and Kompare. It is evident from the result that some factors have impact on duplicate bug reports and some factors do not seem to impact duplicate reports. The factors that impact the duplicate bug report in bug tracking system are number of submitters and size of bug repository. On the other hand project size, project life span and number of developer does not seem to be the factors that impact duplicate bug reports.

Keywords- Open Source Software, Bug Tracking System, Bug Report, Duplicate Bug Report, software Maintenance

I. INTRODUCTION

Open Source Software (OSS) is computer software with its source code that is freely available. Open source software is distributed under license agreement. This license agreement allows the source code to be viewed, modified and shared by user or organizations. Many corporations, large or small, have shown an interest in growing open source software market. It shows some strong differences with traditional software. Open source software is different from traditional proprietary software or closed software in such a way that

proprietary code is developed in private. Unlike traditional software, OSS is developed by geographically distributed teams through the use of the internet. Team members typically work in a distributed environment. Team members are volunteers rather than employee. After development of software, maintenance is required to meet the desired effects. Maintenance is essential for software to keep it up to date and bug free. Bugs are prevalent and widespread in software system. Software maintenance is a very broad activity that includes error correction, enhancements of capabilities, and deletion of obsolete capabilities. Like development, maintenance of OSS is also coordinated through the internet. Open source software projects uses open bug repository during development and maintenance, so that both developers and user can report bugs that they have found, request useful features, modifications and give suggestions. The use of the bug tracking system to organize maintenance activity is widespread. Allowing users to report bugs, help in fixing of bugs and requesting feature is assumed to improve quality of software overall [1]. Bug tracking system contain large amount of bug information that can give deep understanding into the evolution of software project. To improve the reliability of software system reporters, report bugs by writing bug report in bug tracking system like Bugzilla, Jira, Mantis etc. But OSS development has created new challenges to software maintenance. This process of reporting a bug in open source software project is totally distributed and uncoordinated. In this scenario many reporters could submit same report for same problem [2]. The same report which is submitted by several reports referred to as duplicate report. Identification and handling of duplicate bug report is important issue in OSS, because it can result in overhead for development team. If multiple bug report for same bug is not recognized as duplicate it can result in extra effort for development team [3].

To handle duplicate reports a triager need to manually label these report as duplicate report. Due to the large number of reports present in bug repository, it is challenging for the triager to examine all existing reports to detect duplication [1]. Triager manually handles these reports and links them to master report. This task is very time consuming and have negative impact on maintenance and productivity. The study was carried out to find the factors that impact on duplicate bug reports.

II. LITERATURE SURVEY

J. Lerch and M. Mezini [3] 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.

Y.C. Cavalcanti et al. [4] explained that duplicate bug report entries in bug tracking systems have negative effect on software maintenance and evolution productivity. 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. Their work presented an initial characterization study using data from bug trackers from private and open source projects.

Chengnian Sun et al. explained [5] 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.[6] discussed 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.

Anh Tuan Nguyen et.al [7] 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. This approach has been shown to outperform others several approaches in term of both accuracy and time efficiency.

Yuan Tian et al. [8] concluded 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. 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 [9] 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 same fault. This increases the maintainer's triage time.

III. NEED AND SCOPE OF STUDY

In open source software domain it is necessary to handle duplicate bug reports to maintain the efficiency in development and maintenance of software. Due to the abundance of bug reports present in bug repository, it is challenging for the triager to examine all existing bug reports to detect duplicate bug reports. The duplicate recognition is tedious, which requires lots of time and effort. Thus task of duplicate detection is very time consuming and have bad effect on maintenance and productivity of software. Keeping the importance of this issue in mind an exploratory study of duplicate bugs in bug tracking system was done in order to analyze different factors that can cause duplicate bug reports and their impact on duplicate bug reports.

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

IV. OBJECTIVES OF STUDY

The broad objective of our study is to explore the duplicate bugs in bug repository of bug tracking system and more specific objectives are:

1) To find out impact of bug repository size on duplicate bug reports.
2) To investigate the impact of project size on duplicate bug reports.
3) To study the relationship between life span of project and duplicate bug reports.
4) To determine the impact of developers on duplicate bug reports.
5) To study the impact of number of bug submitters has on duplicate bug reports.

V. PROJECTS AND DATA COLLECTION

In this study, projects from open source domain were considered. Five open source projects and their bug repositories were chosen for study. Project chosen were Kompare, Eclipse BIRT, Thunderbird, Firefox for Android and Mandriva Linux. All project chosen use Bugzilla as their bug tracking system. For research, data was collected from projects using Bugzilla as bug tracking system because

- It is generally used by OSS projects
- Bug information is easily downloadable for analysis and study.

Overview of projects under study is given in the Table I

Table I: Overview of Open Source Projects Selected for Study

S. No	*Projects*	*Category*	*Bug Tracking system*
1	**Eclipse BIRT**	**Eclipse based reporting system**	**Bugzilla**
2	**kompare**	**Graphic viewer**	**Bugzilla**
3	**Thunderbird**	**Email application**	**Bugzilla**
5	**Firefox for Android**	**Web browser for android**	**Bugzilla**
6	**Mandriva Linux**	**Linux distribution**	**Bugzilla**

The study was directed over data which was gathered from bug tracking system of five open source projects given in Table 1. Data collected from bug tracking system is given in following Table II.

Table II: Data Collected for Different Parameters for Five OSS Project's Bug Repository

Projects	*version*	*Total submitters*	*Total duplicate bug submitter*	*duplicate bugs*	*Total bugs*	*Total duplicate bugs*	*Line of Code*	*Project life span*	*Number of developers*	*Duplicate %*
Firefox for Android	**Firefox 27**	**10**	**2**	**3**	**7655**	**1942**	*21343*	*3*	*21*	*25*
	Firefox 26	**17**	**8**	**14**						
	Firefox 25	**25**	**9**	**9**						
	Firefox 24	**52**	**24**	**24**						
	Firefox 23	**53**	**24**	**24**						
Thunderbird	**Version 26**	**3**	**1**	**1**	**26867**	**10438**	**1163073**	**10**	**754**	**39**
	Version 25	**6**	**2**	**2**						
	Version 24	**61**	**31**	**32**						
	Version 23	**8**	**3**	**3**						
	Version 21	**18**	**3**	**3**						
Eclipse BIRT	**Version 1.0.0**	**69**	**13**	**14**	**19634**	**1896**	**2152428**	**8**	**125**	**9**
	Version 1.0.1	**94**	**12**	**22**						
	Version 2.0.0	**222**	**82**	**339**						
	Version 1.0.1	**13**	**0**	**0**						
	Version 2.1.0	**198**	**52**	**107**						
Kompare	**Version 3.4**	**114**	**3**	**3**	**226**	**63**	**6274**	**12**	**96**	**28**
	Version 3.4.1	**4**	**3**	**4**						

	Version 4.0.0	**17**	**10**	**10**						
	Version 4.1.1	**1**	**1**	**1**						
	Version 4.1.2	**3**	**2**	**2**						
Mandriva Linux	**Version 2011**	**167**	**65**	**78**	**35861**	**7569**	**1160446**	**14**	**230**	**21**
	Version 2010.2	**121**	**31**	**32**						
	Version 2010.1	**274**	**10**	**134**						
	Version 2010	**879**	**309**	**402**						
	Version 2009.1	**620**	**357**	**308**						

VI. Results and Analysis

This section provides the analysis of data collected from bug tracking system to achieve different objectives. The broad objective of study was to find out the factors that cause duplicate bug reports in bug tracking system of OSS projects. To find out the factors that causes duplicate bug reports study on different factors namely bug repository size, project size, life span of project, number of developers and number of submitters is given in following sections.

A. To find out impact of bug repository size on duplicate bug reports.

Number of bug reports present in bug repository is the size of bug repository. The study was carried out to find whether the size of repository has impact on the duplication of bug reports in bug repository or not. To see the impact of size of bug repository, total number of bug reports and duplicate bug reports present in bug repository of particular project was calculated and relationship between two was analyzed. This can also be analyzed that whether large repository are more susceptible to duplicate bug report than smaller one. Figure 1 shows the graph between size of bug repository and duplicate bug reports in each project's bug repository.

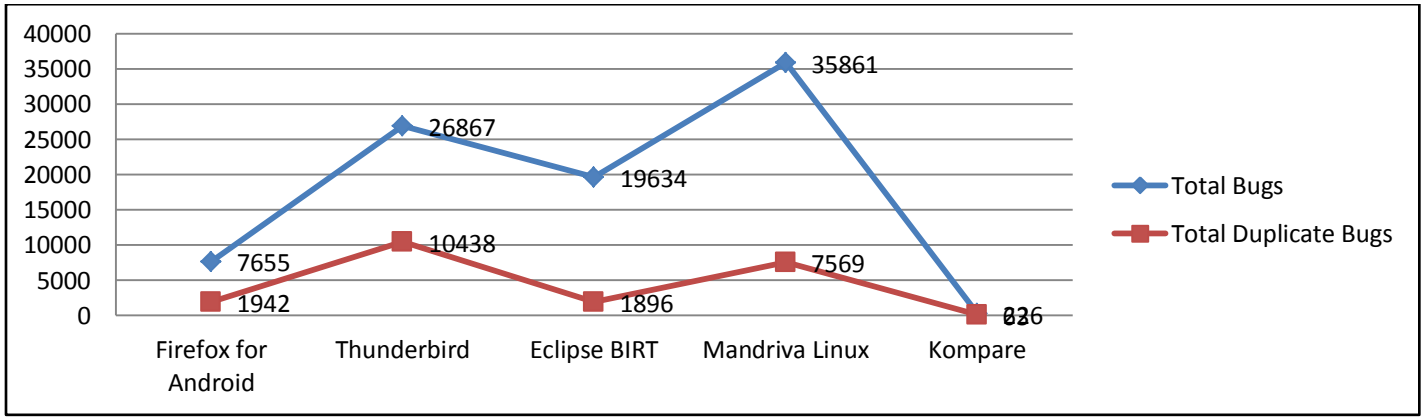


Figure 1: Size of bug repository Vs. Number of duplicate bugs

This is clear from Figure 1 that when size of bug repository increases or decreases duplicate bug reports also increases or decreases for particular project. But it is not necessary that large repository is more susceptible to duplicate bugs that smaller one as Thunderbird has large no. of duplicate than Mandriva linux whereas Mandriva Linux has large bug repository than Thunderbird.

B. To investigate the impact of project size on duplicate bug reports.

Lifespan of software was calculated in year from the time software project has been developed to till date. The study is done to investigate that whether the projects with longer lifespan are more susceptible to duplication of bugs than the project with shorter lifespan. Figure 2 shows the relation between life span of project and duplicate bug reports percentage.

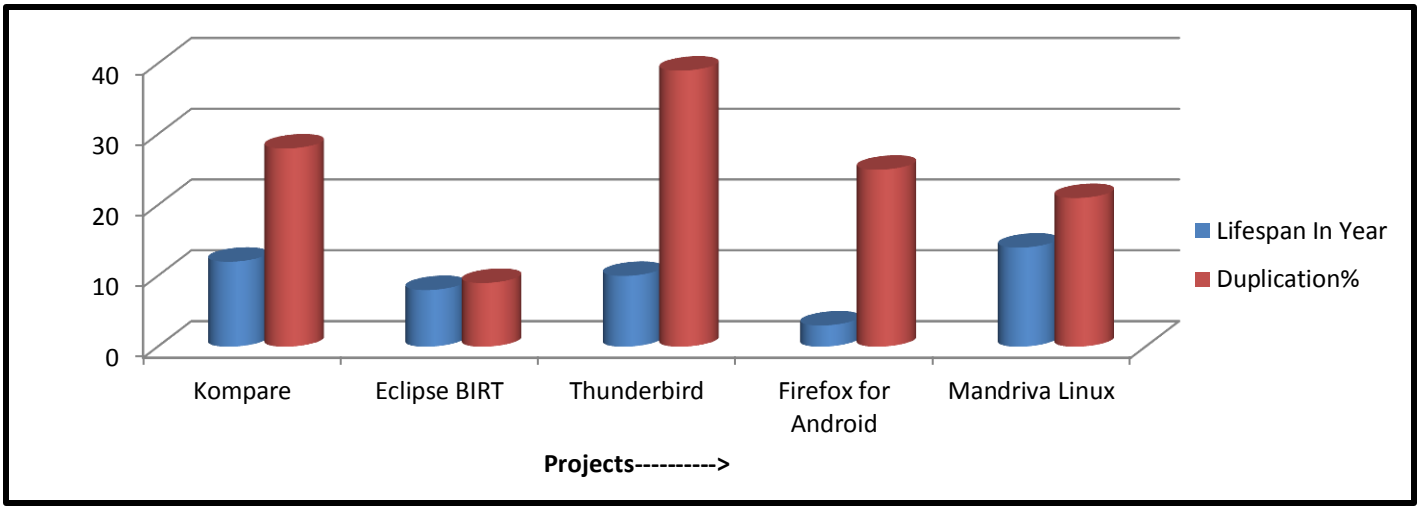


Figure 2: Software Lifespan vs. Duplicate Bug Reports

From Figure 2 it is observed that the lifetime is not a factor that causes duplicate bug reports. Projects with longer lifetimes do not necessarily have more duplicate bug reports than projects with shorter lifetimes. From Figure 2 it is clear that firefox for android have less life span than mandriva linux but more number of duplicates where as eclipse BIRT has less life span than thunderbird and also less duplicates.

C. To study the relationship between life span of project and duplicate bug reports.

Software size was meant in term of Lines of Code (LOC) that software project had. LOC taken is number of line of code at the time when data is collected. Comment and blank lines were discarded, only code lines are taken. The value for all studied project was collected from website http://www.ohloh.net. This was believe that LOC can influence the number of error and leads to more bug report to be submitted and will increase the chance of bug report duplication. For this LOC for all project and duplication % in each project was calculated. Figure 3 shows the relation between LOC and duplicate bug reports.

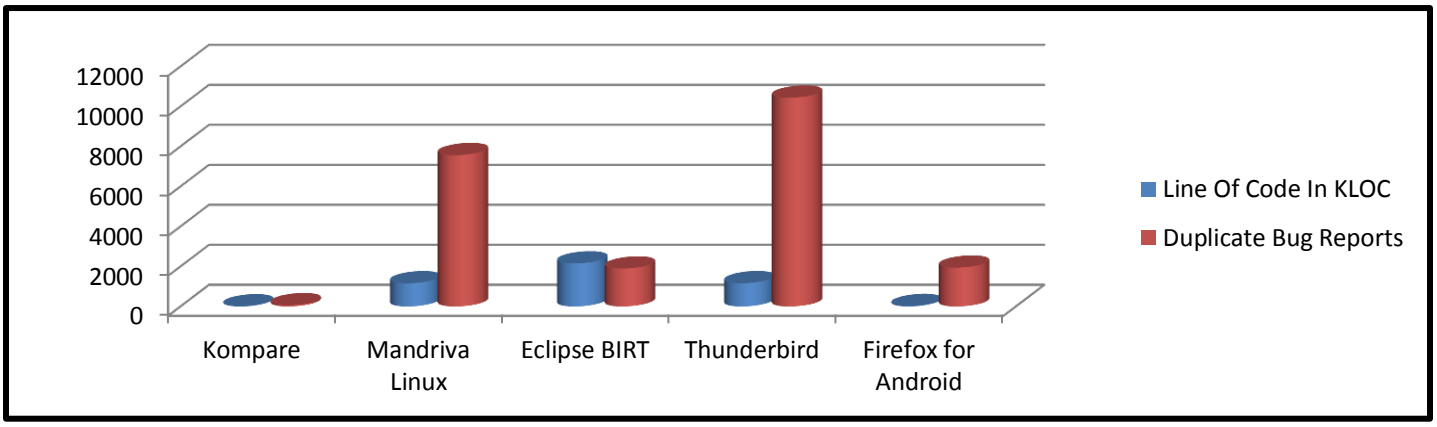


Figure 3: Software Size (LOC) Vs. Duplicate Bug Reports

Figure 3 shows that there is no noticeable pattern between line of code and duplication. So this was concluded that there is no impact of LOC on the duplicate bug. Mandriva Linux and Thunderbird also have approx. equal number of line of code but there is large variation in duplication. This shows that there is no relationship between line of code and bug report duplication in project's bug repository.

D. To determine the impact of developers on duplicate bug reports.

Numbers of developers are the people that are involved in the development of project. This was considered that numbers of developers are equal to the number of people assigned to resolve the bugs. The study was carried out to find whether the number of developers have impact on duplication. For this number of developers and duplicate bug report % in each project calculated and relationship is drawn between two counts which is shown in Figure 4.

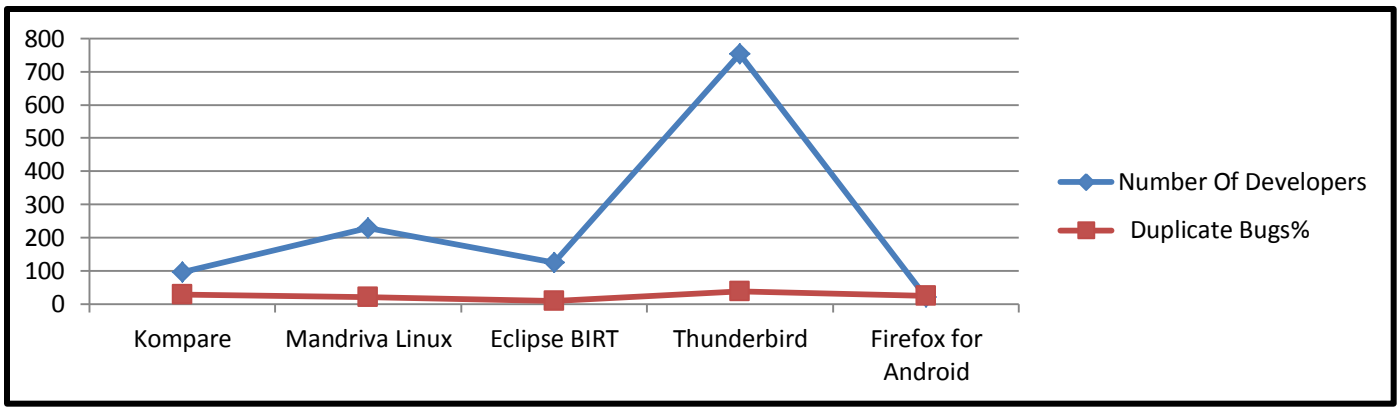
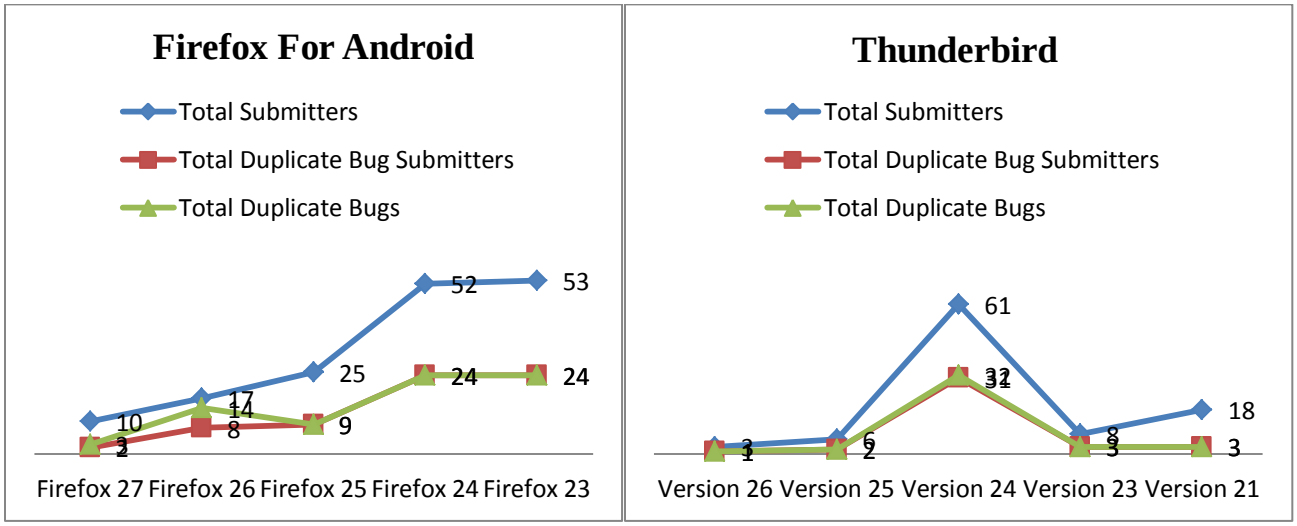


Figure 4: Number of Developers Vs. Duplicate Bug Reports

From Figure 4 it is clear that there is no noticeable pattern that shows some relationship between number of developers and number of duplicate bugs.

E. To study the impact of number of bug submitters has on duplicate bug reports.

Numbers of submitters are the number of people who submitted bug report. Projects were analyzed to find out that whether the number of submitters have impact on duplicate bug present in project. For this total number of submitters, total number of duplicate bug submitters and total number of duplicate bug reports were calculated. The relationship between total number of submitters and duplicate bug submitters, total duplicate bugs and total number of duplicate bug submitters was analyzed. For the study five version of particular project were analyzed. The analysis of objective is drawn in Figure 5.



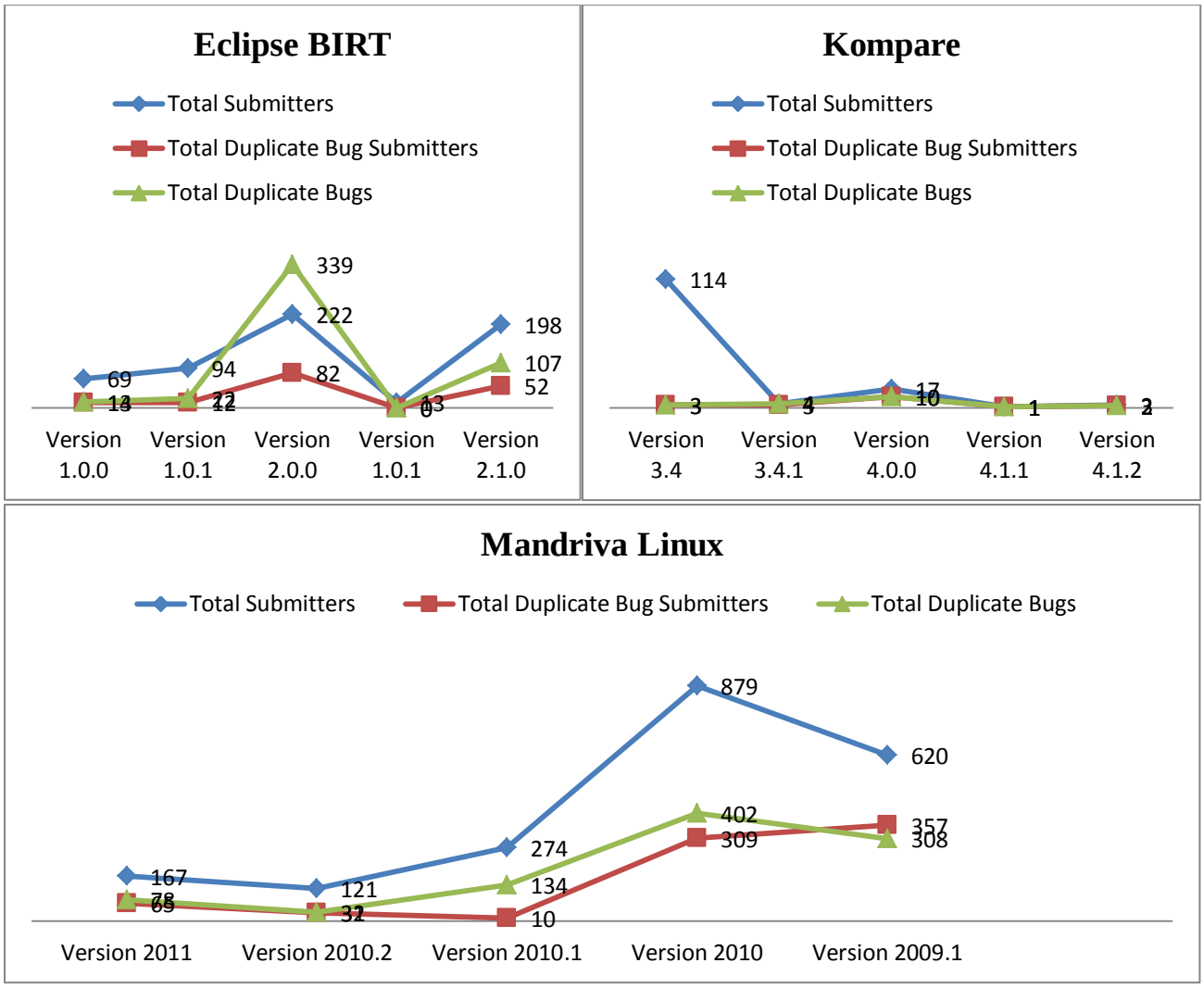


Figure 5: Number of Submitters Vs. Number of Duplicate Bug submitters Vs. Duplicate Bugs

From Figure 5 it is clear that as number of submitter's increases or decrease, total number of duplicate bug submitter also increases or decreases and cause duplicate bug reports for particular version of each project. There is little variation in 2009.1 version of Mandriva Linux where number of duplicate submitters increases and duplicate bugs decreases.

VII. FINDINGS OF STUDY

The following are the findings of study after analyzing the results.

- Bug repository size has impact on duplicate bugs, as bug repository size increases or decreases number of duplicate bug reports also increases or decreases.
- Project lifespan does not seem to be a factor that has impact on duplicate bug reports.
- Software size does not seem as a factor that has impact on duplicate bug reports.
- Number of developers does not seem to be a factor for duplicate bug reports.
- It was observed that number of submitters have impact on the number of duplicate bugs. As number of submitter's increases or decreases number of submitters of duplicate bug also increases or decreases with that number of duplicate bugs also increases or decreases.

VIII. CONCLUSIONS

In OSS domain bug tracking system is a valuable tool that guides the maintenance activity of software. But utility of these systems is hindered by excessive number of duplicate bug reports which is submitted by several reporters for same problem. It is evident from literature survey that identification of these duplicate reports is time consuming and intensifies the already high cost of software

maintenance. The objective of study was to find out the factors that impacts duplication of bug report in bug tracking system. To achieve the objectives exploratory study of bug tracking system was carried out. Five open source software were considered for study. Data collected from bug tracking system of OSS and other websites are total bugs, total duplicate bugs, number of submitters, number of duplicate bug submitters, life span of each project, size of bug repository and project and number of developers. The results of explored data were analyzed to achieve objectives. The factors that impact the duplicate bug report in bug tracking system are number of submitters, size of bug repository. On the other hand project size, project life span and number of developers does not seem to be factors that impacts duplicate bug reports. In future more exploratory studies with more number of open source software and versions can be carried out. Moreover an approach can be presented to handle duplicate bug reports.

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"* , ICSE IEEE, 2008.

[2] Anh Tuan Nguyen, David Lo, Chengnian Sun "*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.

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

[4] 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.

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

[6] Mehdi Amoui, Nilam Kaushik, Abraham Al-Dabbagh, Ladan Tahvildari, Shimin Li and Weining Liu, "*Search-Based Duplicate Defect Detection: An Industrial Experience*" IEEE 2013

[7] 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 (ASE), Proceedings of the IEEE/ACM, 2012.

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

[9] Tomi Prifti, Sean Banerjee and Bojan Cukic, "*Detecting Bug Duplicate Reports Through Local References*", Proceedings of the 7th International Conference on Predictive Models in Software Engineering, ACM,2011.

