Reading Assignment 4

Bhanu Anand, Esha Sharma, Vinay Todalge

i. Reference to the paper:

Nicolas Bettenburg, Rahul Premraj, Thomas Zimmermann, Sunghun Kim.
"Duplicate Bug Reports Considered Harmful . . . Really?" in Software Maintenance, ICSM 2008. IEEE International Conference on Sept. 28 2008 - Oct. 4 2008.

Relation to First Paper:

Authors of First paper have mentioned this paper for stating Duplicate Bugs can be reused in order to have more detailed information about a bug. It had reference to this paper claiming one bug report provide just a partial view of a bug, but multiple bug reports compliments each other.

ii. Important Key-words:

1. Master Report

   Out of many duplicate bug reports, one is selected as root report in order to analyze chain of duplicity. The selected report is called as Master Report.
2. Extended Master Report

   According to theory followed in this paper, authors are not discarding duplicates, and instead they are merging those duplicates with their master report. Merged master report with its following duplicates is called as Extended Master Report.
3. Information Items

   In order to get extended master report, authors have quantified some of predefined fields which are called as Information Items. These fields are important as they support how extended master reports are used to solve problem more efficiently.
4. Triagers

   Traigers are skilled resources who defines duplicates and assigns bugs to respective developers.

iii. Brief Note:

1. *Motivation*:

   Bugs reported by highly technician people might get marked as duplicate of a bugs that is reported as less technical person. Hence there is chance of losing valuable information about bug. In addition to that, instead of simply discarding duplicates, if all duplicates are merged and an extended master report is generated, then it might help to have thorough view on bugs.

2. *Related Work*:

   Authors claims that there are numerous automated triaging techniques which can benefit from extended master reports. They also found evidence that information from bug duplicates can improve existing techniques to automatically assign developers to bug reports.

3. *Study Instruments*:

Authors have made various observations while preparing datasets for their experiments. They claims those as Reasons for duplicate bugs. Some of reasons are '*Multiple failure, one defect'*, '*Intentionally resubmission', Accidental resubmission', 'Duplicate submission at the same time', 'Lazy and unexperience users'.*

4. *Data:*

For experiments in this paper, bug database of Eclipse is used. It consists of 211,843 bug reports, which have been reported between October 10, 2001, and December 14, 2007. Out of these, 16,511 are master reports with 27,838 duplicate reports. They used the UNIX tool "file", which returns the mime-type of a file. Whenever the mime-type of a file starts with "image/", they consider it to be a screenshot. Regular expressions were defined to define patch submission.

5. *Results:*

Experiments were carried with 2 machine learners namely SVM (Support Vector Machine) and Naïve Bayes. Accuracy is considered as measuring Top 1, Top 3 and Top 5 predictions. Top 5 measure delivers highest accuracy for both models. Overall SVM performed better than Naïve Bayes. SVM reached highest accuracy of approximately 65%.

6. *Anti-patterns/Negative results:*

While claiming this theory of merging duplicates, authors have stated some of the negative results as well. Authors are afraid of Results that may not generalize other projects. Sometime addition information slows down the process. It may be harmful and disturbing.

7. *Future Work*:

There are some of the recommendation in the paper for better bug tracking system, some of those are 'providing a feature for efficient merging so that all relevant data is stored at one report', 'renewal of long living bug reports', 'improve pre search for bug reports to avoid submission of duplicates', 'search for master report and add additional information'.
These feature has tremendous potential to explore before they gets into stable situation.

iv. Ways paper would be improved:

1. Authors have not performed 'stemming' (process of handling stop words like 'the', 'a', 'an' etc) thinking of its little benefits. But it does slow down process of finding duplicates in very very large databases. Stop words needs to gotten rid of in order to have efficient processing of duplicate bugs.

2. There is no mention of how automated triaging process would handle biased bug assignment. As automated triager might keep on assigning bugs to same person over some period of time.

3. User/Developers bug filing practices were not taken into consideration while determining *information events* which plays vital role in creating extended master report.