Reading Assignment 2

Bhanu Anand, Esha Sharma, Vinay Todalge

i. Reference to the paper:
Authors: Nicholas Jalbert and Westly Weimer
Paper Name: Automated Duplicate Detection for Bug Tracking Systems
In Proceedings of the International Conference on Dependable Systems & Networks: Anchorage, Alaska, June 24-27 2008.
Relationship to first paper: The authors of the first paper quote the current paper in their related work section. The first paper builds on the material presented in the current paper and proposes a model to predict duplicate reports more accurately using features other than just the textual features.

ii. Important Keywords :

1. Bug Triaging: This is the process of deciding when a given bug should be fixed. It involves deciding in which release a bug should be fixed.
2. False Positive: A bug report which is fixed but gets incorrectly classified as an unfixed bug report is called a false positive.
3. Document Similarity: This is a method which measures the similarity between two documents in the dataset. This is measured by calculating the vector distance between two documents based on the number of words which are common to both.
4. Graph Clustering Algorithms: These algorithms divide a set of input data into clusters wherein each element in a cluster is similar to each other element in that cluster.

iii. Brief Notes :

1. Motivational Statements: Bug Tracking systems are important to guide the maintenance activity of a software. Bug Tracking systems make use of bug reports. The authors found that a significant fraction of bug reports found in bug tracking systems are duplicates. Evaluating these duplicate reports wastes developer time and effort. The authors were hence motivated to find a way to reduce the number of duplicate reports in a system.
2. Related Work: The authors talk about work related to defect report quality in this section. The authors have mentioned a previous research they have done in which they predicted when a bug gets triaged. The authors quote their previous work to compare the appearance of false positives with the current work. The authors also mention a research where in bug reports were assigned to a developer so that the appropriate developer could be involved in a given code. They also talk about previous studies wherein duplicate detection was studied and techniques to mitigate it were investigated. The authors have mentioned a few more studies which are regarding detecting bugs and duplicates and have contrasted their work and the performance of their work with those studies.
3. Data Used: The authors have based their studies on 29,000 bug reports generated from the Mozilla project. These reports span an eight month period from February 2005 to

October 2005. The project contains data from a variety of sources including web browsers, mail clients, calendar applications, and issue tracking systems. The only thing common amongst all these systems was that the same bug tracking system was used in them. This bug tracking system has been employed in the current study.

4. New Results: The authors have proposed a model using which we can identify whether or not a newly submitted report is a duplicate of a previously submitted report. The author uses features of reports already in the system to identify whether or not a newly submitted report is a duplicate or not. The authors propose doing textual analysis of two given reports to predict whether or not the two reports are duplicates.

iv. Three ways the paper could be improved:

1. The authors have only used textual analysis to separate and identify two reports as duplicates. This may not be an exhaustive criteria to identify duplicates. For example, there may be two reports which are duplicates but contain images instead of text where the proposed model will fail.
2. The authors have used only data from one source to apply their study to and evaluate their study. The source studied is open source. The authors could have drawn on data from another source to make their case stronger. Also, they could have used data from at least one corporate source.
3. The model may generate false positives in case there two reports generated very close and which report distinct though very similar bugs as the model is based on identifying textual similarities in bug reports and duplicates are detected based on the extent of this similarity .

v. Reference: Automated Duplicate Detection for Bug Tracking Systems, Nicholas Jalbert and Westly Weimer