Reading Assignment 5

Bhanu Anand, Esha Sharma, Vinay Todalge

i. Reference to the paper:
Chengnian Sun, David Lo, Xiaoyin Wang, Jing Jiang, Siau-Cheng Khoo
"A Discriminative Model Approach for Accurate Duplicate Bug Report Retrieval" in Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering 2010.

Relation to First Paper:

First paper refers this paper for fact that this paper uses only natural language text of bug reports to produce ranking of related bug reports. Also, first paper addresses that fact that some of the 54 features used in this paper are having more importance than other and should be considered accordingly.

ii. Important Key-words:

1. Information Retrieval

   Information Retrieval is method to extract useful information from unstructured documents, most of which are expressed in natural language.
2. Recall Rate

   Recall rate can be defined as the percentage of duplicates whose masters are successfully retrieved in the list.
3. Term Weighting

   For getting similarity between two bug reports, terms used in bug reports are counted and processed accordingly. Term weighting gives a way, by which two bug reports can be compared to each other using words used in describing bugs.
4. Pre-processing

   Before words in bug reports are used for term weighting, there has to be pre-processing on it. It involves tokenization, stemming and stop word removal. This phase is important as it removes noise and speed up further processing.

iii. Brief Note:

1. *Motivation*:

   Bug repositories have variety of bugs. Due to complexity of finding duplicates, it needs lot of efforts and resources as well. In addition to that, not all duplicates state the same information, sometimes they compliments each other and can be used together to get detailed information about bug.

2. *Related Work*:

   Other studies in this area involves less number of features used for finding similarities. Authors have used such 54 features, most of any till at that date, turning more efficient duplicate bug detection. Authors had come up with new way of automatically assigning optimum weight to each feature.

3. *Study Instruments*:
   Authors have made various observations about how model evolves. As new bug are reported, model is trained with new updated training data. They divided these updates in 2 type, Light updates and regular updates. Light updates are made when new bug is far away from master in term of similarities and only *idf* is updated and model is retrained. Regular update are made when a bug finds master then *idf* is updated and new training data is generated and model is trained accordingly. *Idf* is inverse document frequency which tells importance of word towards a document in collection of document.

4. *Data:*
   Authors have used bug report set submitted to Open Office in year 2008 (including 12,732 bug reports), the bug report set of Eclipse in year 2008 (including 44,652 bug reports) to evaluate approach. To evaluate training based approach in long run, bug report set of Firefox (including 47,704 bug reports submitted since Firefox was started in 2002) before June 2007 is used.

5. *Results:*
   Authors continued with bucket based retrieval, where a bucket is considered as one master bug and rest of the bugs are duplicates of that master in that bucket. They have 17–31% relative improvement in Open Office dataset, 22–26% in Firefox dataset and 35–43% in Eclipse dataset.

6. *Future Work*:
   An interesting aspects mentioned is to incorporate threads/discussions on same bug as well on while maintaining duplicate bug reports. This helps to gain some of useful information and larger view on the bug. In addition to that, pattern based classification might be used to extract similarity features.

iv. Ways paper would be improved:

1. As per claim, multiple duplicates compliments each other, authors are not considering the most relevant bug report as master report. They are always marking new bug report as duplicate. There should have been a consideration to all relevant report while deciding master report.

2. Authors have handled only one category of duplicates i.e. duplicates describing same failure. But they did notice and mention that other category might be there in abundant numbers i.e. duplicates showing different failure but originating from same root cause.

3. Users/Developers bug filing practices were not taken into consideration while determining *special features* for comparisons which plays vital role in determining duplicates.