READING ASSIGNMENT VII

DUPLICATE BUG REPORT DETECTION WITH A COMBINATION OF INFORMATION RETRIEVAL AND TOPIC MODELING

Anh Tuan Nguyen, Tung Thanh Nyugen, Tien N. Nyugen, David Lo, Chengian Sun

II. IMPORTANT KEYWORDS

ii1. Duplicate Bug Reports

Report of a bug in software development process for which a bug already exists or another bug exists which has same root cause.

ii2. Natural Language Processing

Branch of computer science, artificial intelligence and computational linguistics which deals with interactions between computers and human languages.

ii3. Vector Space Model

Algebraic model for representing text documents as vectors of identifiers such as index terms which can be used for work such as information filtering, information retrieval, file rankings, etc.

ii4. Machine Learning

Study of pattern recognition and computational learning theory in artificial intelligence which explores study and construction of algorithms that can learn and make predictions on data.

III. BRIEF NOTES

iii1. Motivation

People with different perspective and experience tend to file same bug with different kind of data about the bug. Detection of duplicate bug reports would have benefits in software maintenance in terms of using information in all the available bugs and its duplicates to solve a problem and also save time to solve duplicate bugs.

iii2. Related Work

There have been several approaches to support automatic detection of duplicate bugs. The Jang et al.[1] included methods like traditional information retrieval, Vector Space Modeling based on textual data to solve this problem. Using NLP to find the relevant execution traces to identify duplicate bugs have also been topic of interest to present researchers.

iii3.Hypotheses

This study is expected to help a lot in software industry to reduce software turnaround time when duplicate bugs occurs. This could also help in bug localization techniques by improving the ranking of meta data associated with a bug.

iii4.Baseline Results

The results of this experiment can be compared against results that have been used in studies before like finding duplicate bugs using topic based approach or Support vector using textual words.

IV. IMPROVEMENTS

iv1. All the experiments have been carried on a very high version of physical machines. This tools is expected to run on a developer's system which would typically be a virtual machine and so results running the tools in virtual machines makes sense since this tools looks to be a CPU intensive tool.

iv2. The study would have made more sense if the study was carried out to combine two of the three IR, SVM and NLP techniques before combining all of three together.

Relation to Original Paper

This paper cites original paper where the author have cited the results in topic based bug localization to find duplicate bugs.

References:-

1. *Towards more accurate retrieval of duplicate bug reports - C. Sun, D. Lo, S.-C. Khoo, and J. Jiang - ASE'11*

2. *A Topic-based Approach for Narrowing the Search Space of Buggy Files from a Bug Report - A. T. Nguyen, T. T. Nguyen, J. Al-Kofahi, H. V. Nguyen, and T. N. Nguyen - In ASE'11.*