Reading Assignment 8

Bhanu Anand, Esha Sharma, Vinay Todalge

i. Reference to the paper:

Anh Tuan Nguyen, Tung Thanh Nguyen, Tien N. Nguyen, David Lo, Chengnian Sun "Duplicate bug report detection with a combination of information retrieval and topic modeling" in Automated Software Engineering (ASE), 2012 Proceedings of the 27th IEEE/ACM International Conference 3-7 September, 2012.

Relation to First Paper:

This paper refers to our first paper for comparing result.

ii. Important Key-words:

1. Topic Model

   This is novel model suggested in paper which is designed to address textual dissimilarity between duplicate reports.

2. IR based features

   Apart from the help of topic model, authors are considering textual similarity between two duplicate bugs for duplicates detection. For calculating textual similarity, IR based featured are used to read textual representation of bugs.

3. DBTM

   A DBTM (Duplicate Bug report Detection Model) is new technique suggested in this paper with the help of topic model, which takes advantage of both IR based features and topic based features.

4. Training Dataset

   In this new approach, authors are training their new model with historical training datasets. Authors have maintained collection of technical term/words as Vocabulary as training dataset.

iii. Brief Note:

1. *Motivation*:

   Authors have compared some of duplicate bugs, they found that Duplicate reports describe the same technical issue. Moreover those observations suggest that the detection of duplicate bug reports could rely not only on the technical terms, but also on the technical topics in the reports. Hence there is need of consideration of technical topic as well while detecting duplicate bugs.

2. *Related Work*:

   There has been lot of work on comparing two bug reports based on their textual similarity. Textual similarity is extracted using various model such that BM25F, SVM etc. But no such methods uses technical topic which author claims to use in their novel model.

3. *Hypothesis*:

   As topics are one of many aspects of given system, bugs are mapped against topic for finding duplicates. This helps us to filter bugs accordingly so that bug assignment task can be easily accomplished.

4. *Sampling Procedure*:

   Each aspect/feature of software/system is considered as Topic, represented by words or terms. Bug reports are those reports which shows wrong implementation of such topics. Terms/words used in topics are maintained in one Vocabulary (Voc). Authors have used LDA (A Generative Machine Learning Model), which analyses a bug report as an instance of machine state. LDA can be trained on timely basic with the help of vocabulary maintained with model.

5. *Data:*

   For experiments in this paper, author have used same datasets used in our first paper. They have used 3 large software bug repositories from Mozilla, Eclipse and Open Office, enough variant data to check wide variety of bugs.

6. *Results:*

   Being trained with historical vocabulary, authors claimed that DBTM can improve accuracy up to 20% in detecting duplicate bugs.

iv. Ways paper would be improved:

1. Bug filing practices were not taken into consideration.

2. Only 3 project datasets were used in experiments. For generic validation, this model should be tested against variety of bugs.

3. As claimed, technical topics are also important in their new model, but they have not used those topics while training model. Only vocabulary maintained with respect to duplicate bug is used as training data set. One assumption here could be, overlapping of many terms. But in order to have more generic model, all possible training data sets should be used for training model.

4. Bug assignment can be done once a report is marked as non-duplicate as topic based model already knows what aspect of system a particular bug corresponds to.