Abstract

- Software regression bugs are defects which occur, when a previously working software feature or functionality stops behaving as intended.
- Lot of times code changes or system patches cause an existing feature to misbehave or fail completely.
- Regression bugs are considered to be inevitable and truism in large and complex software systems.
- We present a character n-gram based predictive model, Sarathi for predicting bug inducing change for given regression bug report.

Proposed Approach

- Our proposed approach consists of 3 phases:

  Phase 1: Ground Truth Dataset Establishment

  Phase 2: Feature Extraction

  - Features extracted are of 2 types:
    - Temporal - Time Difference (N) in days between the date of report of the issue and the commit date of the bug inducing revision.
    - Textual - 4 textual features identified using character n-gram approach are:
      1. Similarity between Title of a bug report and the Log Message of the revision.
      2. Similarity between the Description of a bug report and Log Message of the revision.
      3. Similarity between the Cr and Area labels of the issue and the Changed Paths in the VCS.
      4. Similarity between the Title of the issue and the Changed Paths in the VCS.

  Phase 3: Predictive Model

Research Motivation and Aim

“Regression bugs are considered to be inevitable and truism in large and complex software systems.”

- In Google Chromium ITS data (downloaded) 51% of labeled bugs are regression and majority of regression bugs have high priority.

Dataset

- Suitable value of N - Number of look-back days

Results

- 78% of the bugs are reported within 20 days of the induction of the bug.
- Accuracy is maximum i.e. 60% for N=20 and K=75
- Character n-gram based approach for finding textual similarity between issue reports and log messages, files modified can be used for finding culprit revisions.
- The approach can further be improved by improving the pre-processing stage.

Conclusion

- Human intensive task
- No record of previous bug inducing changes
- Level of Detail provided in bug report and log messages
- Noisy data

Technical Challenges

- Noisy data

References

- Alexander Tarvo. Using statistical models to predict software regressions. In ISSRE ’08, pages 259-264. IEEE.
- S. Lal and A. Sureka. A static technique for fault localization using character n-gram based information retrieval model, ISEC 2012