READING ASSIGNMENT II

Bug localization using latent Dirichlet allocation

Stacy K. Lukins, Nicholas A. Kraft, Letha H. Etzkorn

IMPORTANT KEYWORDS

ii1. Information Retrieval(IR) Search Models

Searches based on metadata or content based indexing to obtain relevant data for an information from a collection of information sources.

ii2. Static and Dynamic Bug Localization Techniques

Static Bug localization Techniques gather information for bug localization from source code whereas Dynamic localization techniques gather them from execution traces.

ii3. Probabilistic Latent Semantic Indexing(PLSI)

It is a statistical technique for analysis co-occurrence data which leads to derivation of low dimensional representation of observed variables with respect to their affinity to certain hidden variables.

ii4. Hyper-parameter

Parameter of prior distribution, i.e. probability distribution that would express one's belief about the quantity under study before some evidence is taken into account.

BRIEF NOTES

iii1. Motivational Statements

Generative statistical model which has significant advantage over Information retrieval(IR) models in terms of modularity and scalability. Latent Dirichlet Allocation(LDA), a generative statistical model, has also been shown to be effective in topic based information search and so the authors have applied LDA to automatic bug localization using static technique to evaluate its effectiveness.

iii2. Hypothesis

The authors believe that since the size and complexity of modern software keeps on increasing with new added functionality, using LDA method for automatic bug localization would have save a lot of efforts in maintenance of software products and also improve the bug fixing time and efforts.

iii3. Commentary

The paper uses LDA based techniques on the five different examples on which information based search techniques have been used shows improvements by using LDA. The paper also check if the application of LDA is affected by scalability of software and illustrates that LDA works equally well with scalability.

iii4. Future Work

This model could be to chose a method to select the best topics to be used in LDA approach. One method could be to retrieve results using the LDA approach and then select top 'n' results based on some threshold. Then create a new models with these 'n' topics and check the result.

IMPROVEMENTS

iv1. The paper could have been improved by using the search to take historical searches into account and identify the rank of buggy files based on their count they have identifies as buggy.

iv2. None of the models have taken into account to prioritize the recently changes files in the source code. The recently changed file has higher probability to be buggy.

iv3. Model could also be improved by taking into account that similar bugs might be related to similar fixes.

REFERENCES

1. A Topic-based Approach for Narrowing the Search Space of Buggy Files from a Bug Report - Anh Tuan Nguyen, Tung Thanh Nguyen, Jafar Al-Kofahi, Hung Viet Nguyen, Tien N. Nguyen, Electrical and Computer Engineering Department, Iowa State University