READING ASSIGNMENT IX

ENABLING IMPROVED IR-BASED FEATURE LOCATION

David Binkley, Dawn Lawrie, Christopher Uehlinger, Deniel Heinz

II. IMPORTANT KEYWORDS

ii1. Data Preprocessing

A data mining technique to convert raw data into meaningful data that can be used for study.

ii2. Empirical Study

Phenomenon that is studied through observation and experimental studies.

ii3. Retrieval Models

Model that provides means of identifying documents likely to be of interest given a search query.

ii4. Smoothing

A statistical process to remove noise, random data from a given set of data. This can be ways like taking average, etc.

III. BRIEF NOTES

iii1. Motivation

Modern software development requires a significant support of effective tools which is used to extract information and techniques to aid software engineer. Recent software engineering problems have chosen information retrieval(IR) as a technique for localization in a vast document space. Every IR model involves choosing the right model and selecting the right query. This paper studies impact of (i) query preprocessing, (ii) scraping techniques for query selection, and (iii) the performance impact of choosing the right model. This study can help improve the selection of query and model in IR and hence improve its accuarcy.

iii2. Related Work

There is a considerable body of work related to feature location before this study. Dit.et al.[1] have worked on systematic survey of entire data set along nine key dimensions which includes three dominant types dynamic, textual and static. They have noted that with textual analysis, pre-processing options like stemming have been used considerably but their effects on scope of pre-processing is not studied. Weise et al.[2] and Hill et.al[3] have worked to look at accuracy of pre-processing effects like stemming on the dataset. Marcus et al. have studied impact of choosing different model types.

iii3.Study Instruments

SEMERU dataset used by Dit.et al[1] has been used as the source data for executing this study. Six retrieval models Vector Space Model(VSM), Latent Symantic Indexing(LSI), weighted-sum VSM, and three Query likelihood models Dirichlet-smoothing QL, linear smoothing QL, and latent Dirichlet allocation QL have been used for study. Since this is an investigative study, all the models have been against the dataset.

iii4.Hypothesis

This results for this paper can be used in designing and choosing query for IR based modeling of feature localization. This paper has taken into account the cost of different data pre-processing techniques, so the modeler knows the impact of choosing one technique over another.

IV. IMPROVEMENTS

iv1. The paper has taken precision into account for ranking the results, results for recall would make more sense to also study the inaccuracy of the model

Relation to Original Paper

This paper uses the techniques of bug localization used in original paper and studies impact of query pre-processing and model selection to improve the accuracy of the model using feature location as an example.

References:-

1. *Can better identifier splitting techniques help feature location? - Dit B., Guerrouj L.,Poshyvanyk D., Antoniol G.,2011.*

2. *A comparison of stemmers on source code identifiers for software search - Wiese,A.,Ho,V.,Hill,E.,2011.*

3. *Mining source code to automatically split identifiers for software analysis - Enslen E. Hill, Pollock L.,Vijay Shanker K., 2009.*