Towards a Taxonomy of Approaches for Mining of Source Code Repositories

Huzefa H. Kagdi, Michael L. Collard, Jonathan I. Maletic

Software Development Laboratory <SDML>

Department of Computer Science

Kent State University

Kent Ohio, USA

Motivation

- A number of approaches have been proposed to derive and express changes from source code repositories in a more source-code “aware” manner

- We need better insight of the current research in the MSR community in order to facilitate building efficient and effective MSR tools

Building a Taxonomy

- Draw similarities and variations between six MSR approaches based on three dimensions
  - Entity type and granularity
  - How changes are expressed and defined
  - Type of MSR question

- Define notations to describe MSR to facilitate a taxonomic description of approaches

An Initial Taxonomy

	Entity	Change	Question
Annotation Analysis			
Gall et al	class	syntax and semantic -hidden dependencies	market basket and prevalence
German	file & comment	syntax and semantic - file coupling	market basket and prevalence
Heuristic			
Hassan et al	function & variable	syntax and semantic -dependencies	market basket
Data Mining (association rule)			
Zimmerman et al	class & method	syntax and semantic - association rules	market basket
Differencing			
Raghavan et al	logical statement	syntax and semantic - move	prevalence
Collard et al	logical statement	syntax - add, delete, modify	prevalence

Conclusions

- Most of the approaches except *Differencing* work with fairly high-level entities
- Very different semantic information being is used in these approaches
- Further investigation is necessary to discern between how changes are expressed