Parsing and Modeling Software Engineering Artifacts

Andrea Mocci[1]

Abstract

Artifacts containing natural language, like development emails, tutorials, and Q&A websites (e.g., Stack Overflow), are essential in the practice of software development, and thus they have become a popular subject for software engineering research.

The analysis of such artifacts is particularly challenging because of their heterogeneity: These resources consist of natural language interleaved with fragments of multiple programming and markup languages. Moreover, often these language are not only interleaved, but included in other ones, like XML fragments in Java strings, or method references in JSON values.

The tutorial is aimed at overcoming this challenge by first discussing the state of the art of methodologies to analyze unstructured data, and their current limitations and shortcomings. Then, it focuses on our efforts towards a systematic approach to model contents of such artifacts. This in turn enables novel holistic analyses that fully exploit their intrinsic heterogeneous nature. We describe the theoretical foundations of our StORMeD framework, how it can be used to extract a full-fledged model of a development artifacts, and how it can be leveraged to construct various types of analyses, such as summarization, the analysis of video tutorial fragments, and the construction of a holistic recommender system.

[1] Universita della Svizzera italiana, Switzerland