Ph.D. Thesis Proposal (2) on formal Language Processing (MAPi)

Pedro Rangel Henriques
LP@CCTC – DI/Universidade do Minho

scholar year 2010/11

Problem Domain Concept location for Program Comprehension

Supervisor: Pedro Rangel Henriques + Maria João Varanda
Keywords: Program Comprehension + Program Understanding + Source Code Analysis + Concept Location + Ontologies

Abstract:
An effective program comprehension is reached when is possible to view and relate what happens when the program is executed, synchronized with its effects in the real world concepts. This enables the interconnection of program's meaning at both problem and program domains.

To sustain this statement we need to develop a tool which provides linked views of both domains. This requires that we use the same knowledge representation for both domains. For that, we propose the use of ontologies. Ontologies are defined by sets of objects, classes, attributes and relations in order to conceptualize domains in a systematic way.

Considering the problem domain ontology of the system, the idea of this thesis is to analyze source code, exploring identifiers, comments and prints in order to infer a semantic relation with the concepts described on the ontology.

For that some techniques like natural language processing (Fry,2008), statistics and information retrieval (Vaclav, 2004) can be used. In particular techniques related with concept location (Vaclav, 2002) should be explored: pattern matching, code instrumentation for source code dynamic analysis (Vaclav, 2008), text mining (Cimiano,2005), graph mining and web search engine to discover sentences meaning (Ratiu, 2010).

The Ph.D. is planned to be divided into the following tasks: analyze and write the state of the art; model the problem domain using an Ontology; characterize the notion of Program Identifiers and associated information, and define their extraction and representation; define a way to map program identifiers into the program domain ontology concepts; design the architecture of a system to automatize that process; implement the system; choose case studies and make the system experimental validation.

Conferences (in ranks A to C, according to ERA2010) where the intermediate or final results should be published: ICPC, SCAM, ICSM, ICSE, VisSoft, PASTE, KDV