## Project Summary

Static analysis of software normally takes into consideration only structural information of the source code. We propose to enhance the static analysis of software by adding information concerning the domain semantics. The semantic information is derived from the comments, documentation, and identifier names associated with the source code. Information retrieval methods will be used to extract and analyze the semantic content of software. Combining this information with existing static analysis methods will result in new and improved analysis tools. These tools will help software engineers better analyze and maintain large software systems.

The proposed research has two main objectives. First, a framework will be developed that will allow the combination of semantic and structural information of the source code to develop new and enhanced static analysis methods. The framework will provide means to compute new measures and metrics that describe the software (e.g., cohesion and coupling). Second, the research will study different information retrieval methods and how they can be used to extract relevant semantic information from the source code. An empirical assessment of the framework and the use of semantic information for static analysis will be undertaken.

Existing measurement and analysis methods will be investigated as well as their usage in conjunction with the analysis of the semantic information. Applications of the framework and the new analysis methods to support maintenance and understanding tasks (e.g., re-modularization, clone detection, re-documentation, separation of concerns, identification of interleaved code, identification of patterns and feature location) will be investigated and comparison with existing tools will be performed.

This research and the resulting framework will provide a platform that will help address the following research questions:

- Does the combination of semantic and structural information improve the results of static analysis? What are the best ways to perform this combination?
- Can part of the domain semantics embedded in source code and documentation be used in a systematic fashion for static analysis?
- Which information retrieval methods work best in extracting meaningful semantic information from the source code and its associated documentation?
- What software measures and metrics can be computed using semantic information?
- What software maintenance tasks are best supported by such analysis methods?