Master Thesis Topic

Empirical Study on Query Expansion Techniques for Semantic Code Search

Motivation and Background
Source code reuse is of great importance for software development. Therefore, code search is a crucial part of developers’ daily activities [1]. To improve the effectiveness and efficiency of code search, one direction is to refine and reformulate poor queries that they could be more syntactically and semantically related to the search target. There exist various approaches [2,3] on query expansion for code search, however, most of them are restricted to a certain search scenario or domain and their effectiveness has not been compared systematically [4].

Goals
The student should build a comprehensive corpus including query-code pairs as the gold set for evaluation. Based on this benchmark, the student should assess and compare the performance of the state-of-the-art approaches for query expansion based on various metrics.

Description of the Task
This task will involve natural language processing, information retrieval, and machine learning techniques. To build such a corpus, data mining skill is also required. A detailed description of the task will be given personally on interest.

Research Type
Theoretical Aspects:  *****
Industrial Relevance:  ****
Implementation  ****

Prerequisite
The student should be enrolled in the master of computer science program, and has completed the required course modules to start a master thesis.

Skills required
Programming skills in (preferably) Java/Python, understanding of, or willingness to learn, the architectural and statistical foundations needed for the project.

References

Contacts
Minxing Tang, Humboldt-Universität zu Berlin, Institut für Informatik, Lehrstuhl Software Engineering, Unter den Linden 6, 10099 Berlin, Germany

Application
Please contact during office hours or write an email with the title: “ESQET-SCS” to se-career@informatik.hu-berlin.de