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

[1] Reiss, Steven P. "Semantics-based code search." *2009 IEEE 31st International Conference on Software Engineering*. IEEE, 2009.

[2] Lv, Fei, et al. "Codehow: Effective code search based on api understanding and extended boolean model (e)." *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 2015.

[3] Liu, Jason, et al. "Neural query expansion for code search." *Proceedings of the 3rd acm sigplan international workshop on machine learning and programming languages*. 2019.

[4] Yan, Shuhan, et al. "Are the Code Snippets What We Are Searching for? A Benchmark and an Empirical Study on Code Search with Natural-Language Queries." *2020 IEEE 27th International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 2020.

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