Семантический поиск программного кода

Мочалов Никита Сергеевич

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: mochalov.n@yandex.ru

Научный руководитель *— Головин Игорь Геннадьевич*

Семантический поиск кода — это поиск программного фрагмента по его словесному описанию на естественном языке. Задача представляет не только академический интерес: такая функциональность может быть встроена в любую поисковую систему, где в документах может встречаться программный код.

Для задач информационного поиска в последние годы успешно применяются нейросетевые методы. В том числе они нашли применение и в задаче семантического поиска программного кода [1].

Проблема нейросетевых методов в том, что им нужны большие наборы данных для обучения, также результаты работы нейронной сети трудно интерпретировать, а поиск ближайших фрагментов кода для запросов работает медленно. Заметим, что полнотекстовый поиск, который, например, реализован во многих СУБД, лишен этих недостатков.

В данной работе представлен новый подход, который был успешно реализован в системе семантического поиска для языка программирования Go. Главная идея - использовать полнотекстовый поиск, но с некоторыми особенностями. Поиск в реализованной системе производится по идентификаторам и комментариям из программных фрагментов. Эти идентификаторы и комментарии дополняются документацией на естественном языке для используемых пакетов и функций. Эта документация извлекается автоматически из системы хранения документации GoDoc [2].

Далее запрос пользователя ищется по данным из базы вопросно-ответного ресурса StackOverflow. Это вопросно-ответный портал для программистов, данные которого доступны для скачивания любому желающему. Поэтому в этой базе можно найти программные признаки, которые близки пользовательскому запросу. Эти признаки используются для переранжирования наилучших результатов из предыдущего шага.

Оценка разработанной системы семантического поиска была произведена на наборе данных CodeSearchNet Corpus [1]. По метрике Mean Reciprocal Rank, наиболее подходящей для поиска фактоло-

гической информации, нейронная сеть архитектуры Neural Bag Of Words [3] получила результат MRR=0.36. В то же время представленная здесь система показала себя лучше: ее оценка MRR=0.56. Эта оценка выглядит обнадёживающе, но еще будет уточнена с привлечением внешней экспертизы [4].

Разработанная система семантического поиска обладает целым рядом достоинств. У нее хорошее качество поиска, ей не нужны наборы данных для обучения, при этом сохраняется интерпретируемость и высокая скорость работы.

Литература

1. Husain H., Wu H. H., Gazit T., Allamanis M., Brockschmidt M. CodeSearchNet Challenge: Evaluating the State of Semantic Code Search // arXiv preprint, 2019, arXiv:1909.09436
2. Система хранения документации для проектов на языке Go `https://godoc.org/`
3. Sheikh I., Illina I., Fohr D., Linares G. Learning Word Importance with the Neural Bag-of-Words Model // ACL, Representation Learning for NLP (Repl4NLP) workshop, 2016, Berlin, Germany
4. Разработанный автором работы сайт для разметки результатов поисковой выдачи `http://gosearch.duckdns.org/`