Московский государственный университет имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра алгоритмических языков

Жуков Павел Николаевич

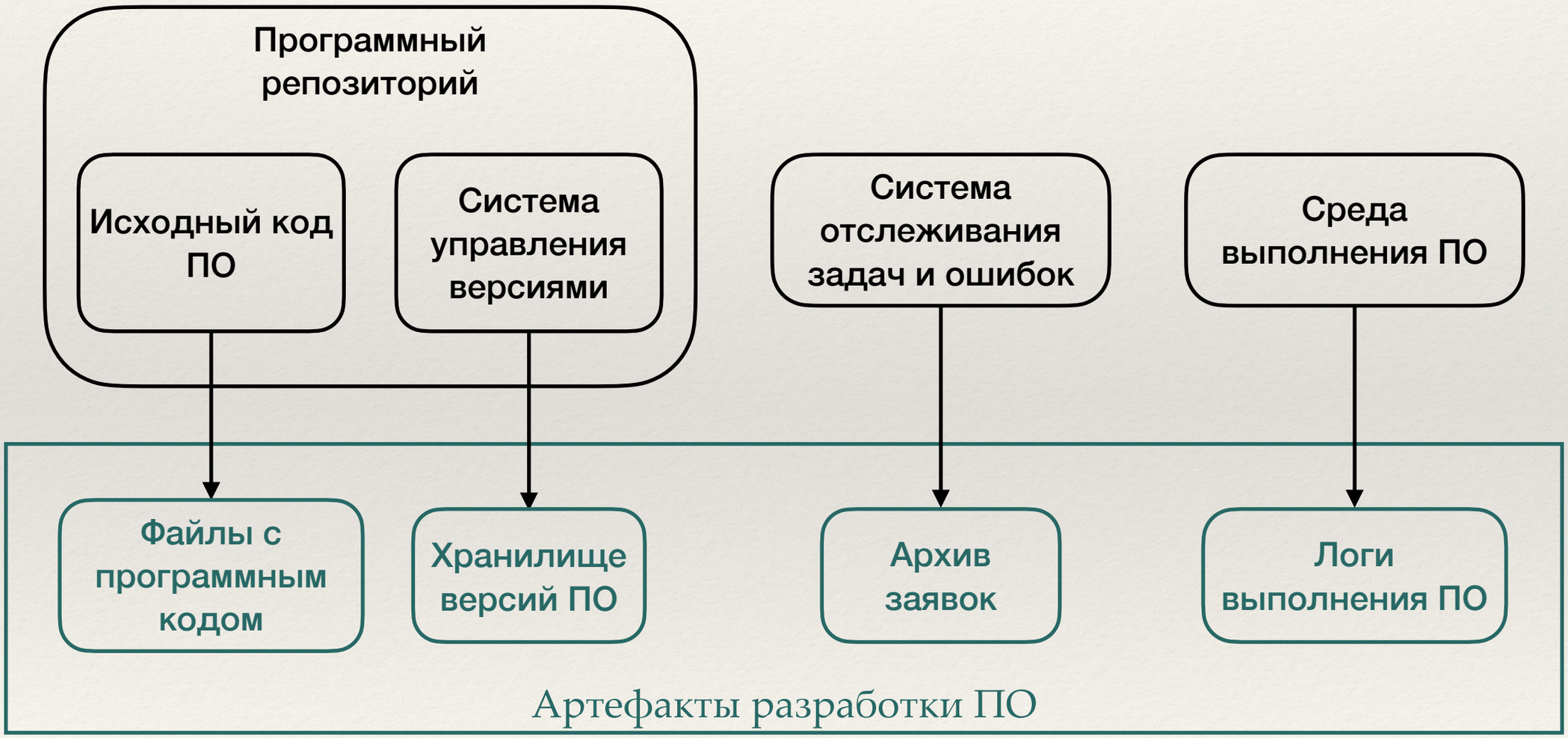
Система поиска информации в программных репозиториях

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Научный руководитель:
к.ф.-м.н., доцент Головин Игорь Геннадьевич

Введение

Разработка программного обеспечения приводит к появлению различного рода артефактов.



Артефакты разработки ПО

Актуальность работы

- Количество вспомогательных средств для разработки растёт, а следовательно, и объём артефактов.

- Существуют специалисты, которые не обязаны понимать код и другие артефакты, но могли бы извлечь из них полезную информацию.

- Не существует программных средств, позволяющих осуществлять поиск по коду, объединённому с его артефактами.

Постановка задачи

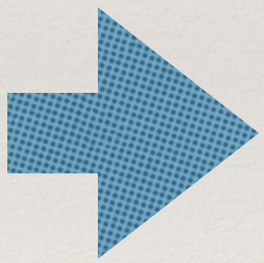
Разработать систему поиска информации в программном репозитории, позволяющую задавать вопросы на естественном языке и использовать информацию из артефактов программного обеспечения, в частности системы контроля версий.

При возможности система должна давать краткий ответ на естественном языке.

Ключевой принцип

Разбить исходный код на методы, объединив каждый из них с релевантной информацией из артефактов, и затем преобразовать в документы, имитирующие текст на естественном языке.

```
void readCommand() {
    try {
        String input = reader.readLine();
        Command command = parser.parse(input);
        if (command != null) {
            commandSender.send(command);
        }
    } catch (IOException e) {
        logger.log(Level.SEVERE, "Exception is thrown", e);
    }
}
```



```
Method "readCommand" was implemented by
Yulia. Its purpose is read command for
input console. The method was created
with message: "Create InputClass.". The
method was created on September 04,
2019. The method tokens are: read line
parse send log.
```

Разработанное решение

Эксперименты

Эксперименты проводились на Java-проекте из 26 файлов, 56 методов, разработанных 10 авторами в 237 коммитах. Для каждого метода запросы составлялись вручную по разным принципам.

	Стали лучшими	В тройке лучших	В пятёрке лучших	В десятке лучших	Не попали в десятку	MRR	Точность BERT
«Who?»	24	20	9	2	1	0,63601	1,0
«Who?» с синонимами	18	28	7	2	1	0,57236	1,0
«When?»	22	23	7	4	0	0,61209	1,0
«When?» с синонимами	17	32	4	3	0	0,57102	1,0
Поиск по одному слову	37	12	6	1	0	0,79153	—

Зависимость ранжирования и точности ответа BERT от типов запросов

Результаты

Основные результаты данной исследовательской работы :

- впервые был предложен подход к построению системы поиска информации в программных репозиториях, позволяющий связать код с артефактами и предоставить возможность получения краткого ответа на вопрос на естественном языке;

- подход был реализован в виде программного средства с использованием современных моделей нейронных сетей BERT и FastText для поиска по проектам на языке Java, и имеет пользовательский веб-интерфейс;

- была показана адекватность реализации, её устойчивость к появлению синонимов в запросах и высокая точность извлечения кратких ответов на вопросы на естественном языке.