РЕЦЕНЗИЯ

на выпускную квалификационную работу обучающегося СПбГУ
Ковалева Дмитрия Александровича *(ФИО)*
по теме «Синтаксический анализ данных, представленных в виде контекстно-свободной грамматики»

Работа Ковалева Дмитрия Александровича посвящена разработке алгоритма синтаксического анализа множества строк, представленных в виде контекстно-свободной грамматики. В то время как большинство методов синтаксического анализа принимают в качестве входа строку, некоторые разбирают регулярный язык, в данной работе предпринята оригинальная попытка расширить область применения на контекстно-свободные грамматики. Тем не менее, задача является в общем случае неразрешимой, поскольку к ней сводится определение непустоты пересечения двух контекстно-свободных языков. Таким образом, было необходимо установить, в каких случаях решение данной задачи является возможным, и предложить соответствующий алгоритм, позволяющий находить данное пересечение. Одной из мотиваций работы является поиск подстрок в геноме, удовлетворяющих определенному шаблону, поскольку как и шаблон, так и геном могут задаваться контекстно-свободной грамматикой, и потому алгоритм должен уметь находить пересечение одного языка и всех возможных подстрок другого языка.

В рамках работы были предложены условия, при которых поставленная задача является разрешимой, и разработан алгоритм, позволяющий находить требуемое пересечение. Тем не менее в работе есть множество теоретических недостатков. Отсутствует доказательство корректности алгоритма, и, следовательно, непонятны его ограничения, что является серьезной проблемой, поскольку из его кода сложно заключить, почему он будет работает на определенных примерах, но не может работать в общем случае. Очевидно, подразумевается, что алгоритм будет работать в предложенных условиях разрешимости задачи, но это также не было отражено в работе. Также алгоритм находит не все подстроки, а только те, которые начинаются и заканчиваются в некотором, одном и том же, правиле.

Кроме того, требуется более тщательное тестирование алгоритма. Во-первых, оно может помочь установить более слабые ограничения на грамматики. Во-вторых, оно может показать более точную зависимость времени от размера входа. Здесь надо отметить, что в работе в качестве размера входа использовался размер грамматики. Хотя он, возможно, важен на практике, на него сильно влияет степень сжатия строки (размер может быть как линейным, так и логарифмическим), и потому зависимость времени разбора от него может сильно варьироваться. Такая метрика, как изначальный размер строки, может являться более предпочтительным выбором для целей тестирования.

В работе было доказано, что исходная задача разрешима для случая, когда одна из грамматик задает конечное множество строк. Поскольку подобные языки являются подмножеством регулярных, для которых задача также разрешима, встает вопрос о применимости данного алгоритма, поскольку пользователь может изначально сжимать данные в виде конечного автомата, тем самым получая возможность обработать более широкое множество. Кроме того, в работе не хватает сравнения разработанного алгоритма и уже существующих для разбора регулярных выражений в случае, когда данные представляют собой конечное множество строк.

Необходимо заметить, что данная работа является одной из первых в данной области, и в дальнейшем возможно как установление точных границ применимости алгоритма, так и его изменение для покрытия более общего случая. Также стоит отметить ясную подачу материала в работе.

Проверка ВКР на предмет наличия/отсутствия неправомерных заимствований показала, что работа неправомерных заимствований не содержит.

С учетом указанных замечаний, считаю, что студент справился с поставленной задачей и заслуживает оценку «хорошо».

«__4_»_____июня____ 2017 г. _____ _Авдюхин Д. А.___

Подпись *ФИО*