Заголовок статьи для Трудов ИСП РАН

1Ю.А. Сусанина <[st049970@student.spbu.ru](mailto:st049970@student.spbu.ru)>

2А.Н. Явейн <yaveyn@yandex.ru>

3С.В. Григорьев <s.v.grigoriev@spbu.ru>

1, 3Санкт-Петербургский государственный университет,

199034, Россия, г. Санкт-Петербург, Университетская наб., д. 7/9.

Аннотация. Аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация аннотация.

**Ключевые слова:** синтаксический анализ; контекстно-свободные грамматики; матричные операции.

**DOI: ???**

**Для цитирования:** Сусанина Ю.А., Явейн А.Н., Григорьев С.В. Заголовок статьи. Труды ИСП РАН, том ?, вып. ?, 2019 г., стр. ?-?. DOI: ???

1. Введение

После того, как Ноам Хомский внес существенный вклад в развитие теории формальных языков, выделенный им тип грамматик — контекстно-свободные (КС) — начали активно изучаться и в дальнейшем нашли широкое применение во многих областях [], прежде всего, в информатике — для описания естественных языков и языков программирования. Также существует множество исследований, которые показывают эффективность использования КС-грамматик в биоинформатике [].

Хорошим примером является возможность применения формальных языков для решения задач распознавания и классификации, некоторые из которых основаны на том, что вторичная структура последовательностей ДНК и РНК содержит в себе важную информацию об организме. Характерные особенности вторичной структуры могут быть описаны с помощью КС-грамматики. Это позволяет свести проблемы распознавания и классификации к задаче синтаксического анализа (определения принадлежности некоторой строки к языку, заданному грамматикой). То есть наличие подпоследовательностей, обладающих некоторыми особенностями, а также их расположение относительно друг друга помогают получить информацию о происхождении организма. Основной задачей было найти алгоритм синтаксического анализа, легко адаптируемый к задаче поиска подстрок, и, кроме того, найденное решение должно быть максимально эффективным, так как такая область применения, как биоинформатика, предполагает работу большими объемами данных.

Большинство алгоритмов синтаксического анализа либо работают за кубическое время (Касами [], Янгер [], Эрли []), либо применяются только к определенным подклассам КС-грамматик (Бернарди, Клауссен []). На данный момент самым асимптотически эффективным алгоритмом синтаксического анализа, работающим с любой КС-грамматикой, является алгоритм Валианта []. Результатом его работы для линейного входа является матрица разбора, каждый элемент которой отвечает за выводимость конкретной подстроки. Валиант смог добиться улучшения вычислительной сложности за счет использования матричных операций. Для входной строки длины n алгоритм заканчивает свою работу за время , где — время, необходимое для перемножения двух булевых матриц размера . Более того, данный данный алгоритм был расширен для конъюнктивных и булевых грамматик, которые обладают наибольшей выразительностью []. Однако алгоритм Валианта плохо применим к описанной выше проблеме поиска подстрок, так как он будет выполнять много лишних вызовов перемножения матриц.

В данной работе предложена модификация алгоритма Валианта. За счет изменения порядка вычисления перемножений матриц появилась возможность разбиения матрицы разбора на слои непересекающихся подматриц. Предложенный подход частично решает проблему поиска подстрок. Кроме того, каждая матрица слоя может обрабатываться независимо, что в дальнейшем позволит повысить эффективность алгоритма, используя параллельные техники.

Работа организована следующим образом. В разделе 2 даны основные понятия и приведен исходный алгоритм Валианта; в разделе 3 представлена модификация данного алгоритма, легко адаптируемая к задаче поиска-подстрок и позволяющая повысить использование параллельных техник, а также доказана корректность и приведена оценка сложности модифицированный версии; в разделе 4 показана применимость предложенного нами подхода к задаче поиска подстрок; в разделе 5 представлены результаты проведенных экспериментов; заключение и направления будущих исследований приведены в разделе 6.

2. Обзор

В этом разделе мы введем основные определения и алгоритм Валианта, на котором основывается предложенная в данной работе модификация. Прежде чем описать алгоритм Валианта, мы покажем, как работают табличные алгоритм синтаксического анализа.

2.1 Терминология

Грамматикой будем называть четверку , – конечное множество терминальных символов, – конечное множество нетерминальных символов, –конечное множество правил вида , где и – стартовый символ.

Грамматика называется контекстно-свободной (КС), если любое ее правило имеет вид , где .

КС-грамматика называется грамматикой в нормальной форме Хомского, если любое ее правило имеет одну из следующих форм:

где , – язык, порождаемый грамматикой .

Как будем обозначать язык, порождаемый грамматикой

2.2 Алгоритм синтаксического анализа, основанный на перемножении матриц

Задачей синтаксического анализа является проверка принадлежности входной строки языку, порождаемому некоторой грамматикой.

Взятый за основу в данной статье алгоритм Валианта относится к табличным методам синтаксического анализа, главная идея которых – построение для входной строки и КС-грамматики в нормальной форме Хомского таблицы (далее матрицы) разбора размера , где

Элементы матрицы должны заполняться последовательно, начиная с диагонали:

Затем будут вычисляться по формуле , где

Входная строка принадлежит языку тогда и только тогда, когда .

Если все элементы данной матрицы заполнять последовательно, то вычислительная сложность данного алгоритма будет равна . Наиболее затратной по времени операцией является вычисление, и Валиант смог реорганизовать порядок заполнения элементов матрицы разбора так, что стало возможным перенести эти вычисления на перемножение булевых матриц.

Входные данные: G – КС-грамматика, , , где степень двойки

main():

accept if and only if

compute:

**if** **then**

;

complete:

**if** **and** **then**

;

**else if** **and** **then**

;

**else if** **then**

;

;

; ;

;

;

;

;

;

;

;

Листинг 1.  Алгоритм Валианта

Listing 1. Valiant’s algorithm

Сначала введем понятие перемножения двух подматриц матрицы разбора .

Пусть – две подматрицы , тогда , где и .

Теперь можно представить как перемножение булевых матриц (для каждой пары нетерминалов). Определим матрицу, соответствующую паре нетерминалов ,как . Тогда тогда и только тогда, когда . Также заметим, что . Более того, каждое из перемножений булевых матриц может быть обработано независимо.

Эти изменения позволили добиться сложности , где — время, необходимое для перемножения двух булевых матриц размера .

Алгоритм Валианта представлен на листинге 1. Все элементы матриц и инициализируются пустыми множествами. Затем эти элементы последовательно заполняются двумя рекурсивными процедурами.

Процедура корректно заполняет все для всех .

Процедура заполняет все для всех иДля корректной работы этой процедуры, во-первых, необходимо, чтобы элементы для всех и , таких что и уже были построены. Во-вторых, текущее значение для всех и , таких что и, должно быть следующим:

Деление на подматрицы во время выполнения этой процедуры показано на рис.1.

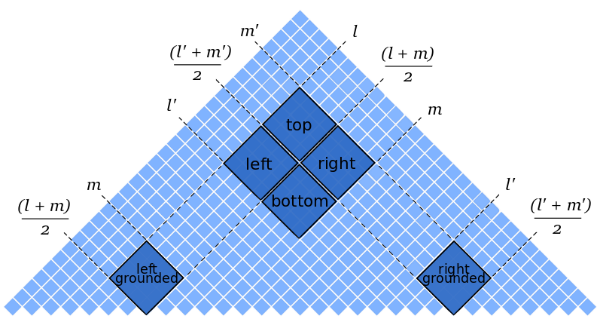


Рис. 1. Деление матриц, использованное в процедуре

Fig. 1. Matrix partition used in procedure

На рис.2 представлен фрагмент работы алгоритма Валианта. Здесь рассмотрено только начало заполнения таблицы , потому что в дальнейшем мы укажем на разницу между исходной версией и нашим подходом.

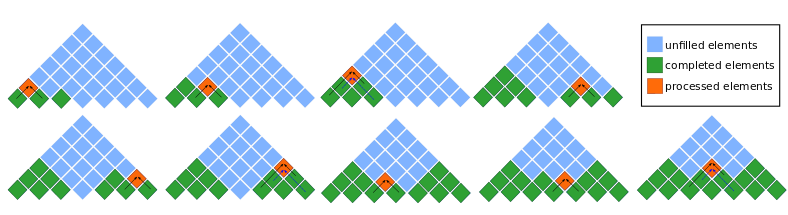


Рис. 2. Пример работы алгоритма Валианта

Fig. 2. An example of Valiant’s algorithm

3. Модификация

В данном разделе мы представляем модификацию алгоритма Валианта. За счет изменения порядка вычисления подматриц в исходном алгоритме модифицированная версия обладает такими практическими преимуществами, как адаптация для решения задачи поиска подстрок и упрощение использования параллельных вычислений.

Главное отличие предложенной модификации – это возможность разделения матрицы разбора на слои непересекающихся подматриц одинакового размера. Пример разбиения матриц на такие слои представлен на рис.3. Каждый слой состоит из квадратных подматриц, размер которых равен степени двойки. Слои заполняются последовательно снизу вверх, и каждая матрица слоя может обрабатываться независимо, что позволяет значительно упростить разработку параллельной версии алгоритма.

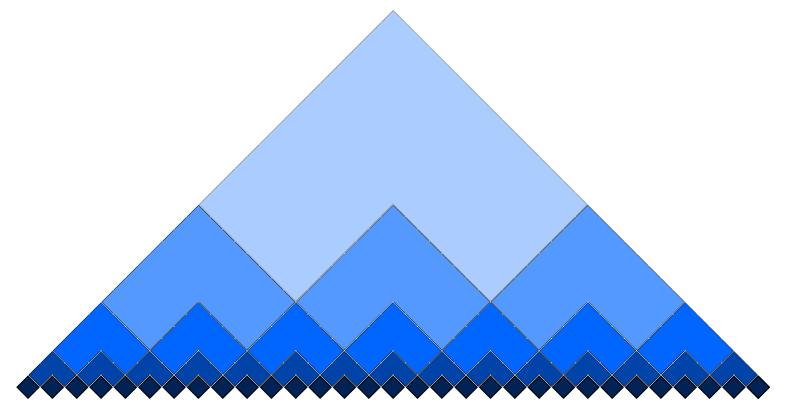


Рис. 3. Деление матриц на V-образные слои

Fig. 3. Matrix partition on V-shaped layers

Простой пример работы модификации показан на рис. 4. Нижний слой (из подматриц размера 1) вычисляется заранее, и заполнение матрицы начинается со второго слоя. (Здесь и далее, под слоем будем понимать некоторое множество подматриц матрицы разбора.)

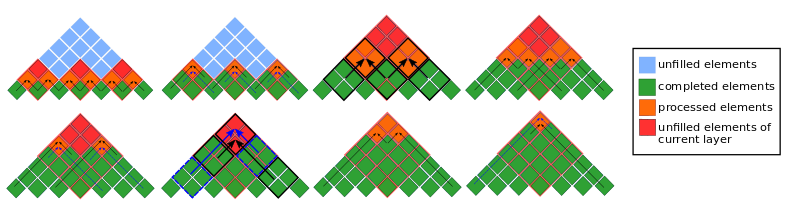


Рис. 4. Пример работы модификации алгоритма Валианта

Fig. 4. An example of the modification of Valiant’s algorithm

Модифицированная версия алгоритма представлена на листинге 1. Процедура заполняет нижний слой матрицы (), а затем разделяет матрицу на слои, так, как было описано ранее, которые корректно вычисляются в процедуре Таким образом, вызов функции заполнит всю матрицу разбора и вернет информацию о выводимости строки для заданной грамматики.

Для краткости, определим несколько дополнительных функций: *, , , ,*  and , которые возвращают подматрицы для матрицы , аналогично разбиению матрицы в алгоритме Валианта, который был представлен на рис. 2.

Также определим функции для слоя подматриц:

Процедура принимает слой непересекающихся подматриц одинакового размера . Для каждой эта процедура вычисляет *, , .* Для корректной работы этой процедуры, во-первых, необходимо, чтобы элементы и для всех и , таких что и уже были построены. Во-вторых, текущее значение для всех и , таких что и, должно быть следующим:

Процедура также принимает на вход набор подматриц , но каждую заполняет полностью. Ограничения на входные данные такие же, как у процедуры *,* за исключением условия на , которое в данном случае не нужно.

Другими словами, вычисляет весь -образный слой , а – это вспомогательная функция, необходимая для вычисления меньших квадратных подматриц слоя *.*

И наконец процедура , где – это массив троек подматриц, представляет основной шаг алгоритма: перемножение матриц. Стоит заметить, что в отличие от исходного алгоритма и каждый может быть выполнен параллельно.

Входные данные: G – КС-грамматика, , , где

main():

**for** **do**

**for** **do**

accept if and only if

constructLayer:

completeLayer:

**if** **then**

**for** **do**

**else**

complete:

;

;

;

;

;

;

performMultiplication():

for do

Листинг 2.  Модификация алгоритма Валианта

Listing 2. Modification of Valiant’s algorithm

Для представленного алгоритма справедливы следующие утверждения.

**Лемма 1.** Пусть – слой. Если для всех :

* для всех и , таких что и ,

* для всех и , таких что  и,

Тогда процедура возвращает корректно заполненные для всех и , таких что идля каждой .

**Доказательство.**

*По индукции по размеру матриц в слое .*

**Теорема 2.** (*Корректность алгоритма*). Алгоритм из листинга 2 корректно заполняет для всех и , и входная строка тогда и только тогда, когда **.**

**Доказательство.**

Перед тем, как доказать утверждение теоремы, докажем по индукции, что все слои матрицы разбора вычисляются корректно.

*База индукции.* Слой размера корректно заполняется в строках 2-3 листинга 2.

*Индукционный переход.* Предположим, что все слои размера вычислены корректно.

Обозначим слой размера как . Будем рассматривать одну матрицу слоя так, как для остальных подматриц их заполнение будет проходить аналогично.

Рассмотрим вызов процедуры .

Заметим, что все для всех и , таких что и *,* уже корректно заполнены, так как эти элементы лежат в слоях, которые уже вычислены по индукционному предположению.

В начале выполнения процедуры , добавляет к каждому все пары нетерминалов , такие что для всех *sublayer(M)* и , такие что для всех . Теперь все предусловия для вызова процедуры выполнены и он вернет корректно заполненные .

Затем процедура вызывается с аргументами и и добавляет пары нетерминалов , такие что и , такие что к каждому для всех . Так как (из построения слоя), условия на матрицу выполнены и вернет корректно заполненный .

Таким образом, возвращает корректные для всех для всех слоев матрицы и строки 4-6 листинга 2 возвращают все , ч. и т.д.

**Лемма 2.** Пусть  **–** это количество вызовов процедуры , где для всех выполнено .

* для всех ;
* для всех матрицы размера перемножаются ровно раз.

**Доказательство.**

Сначала докажем первое утверждение по индукции по .

*База индукции.* При:Следовательно,

*Индукционный переход.* Предположим, что для всех .

Рассмотрим .

Заметим, что функция возвращает матриц размера , то есть в вызове процедуры вернет матриц размера . Также, будет вызвано 3 раза для левых, правых и верхних подматриц матриц размера . Кроме того, вызывается 4 раза для нижних, левых, правых и верхних подматриц матриц размера , за исключением левых, правых и верхних подматриц матриц размера матриц, которые к этому моменту уже были посчитаны.

Таким образом, .

Теперь мы знаем, что , и можем доказать второе утверждение: посчитаем количество перемножений матриц размера . вызывается 3 раза, и . То есть, количество перемножений подматриц размера равно , ч. и т.д.

**Теорема 2.** (*Оценка сложности алгоритма*). Пусть - длина описания грамматики и – длина входной строки. Тогда алгоритм из листинга 2 заполняет матрицу за , где — время, необходимое для перемножения двух булевых матриц размера .

**Доказательство.**

Так как в лемме 2 было показано, что количество перемножений матриц не изменилось по сравнению с исходной версией алгоритма Валианта, то доказательство будет идентично доказательству, приведенному Охотиным [].

Таким образом, мы доказали корректность предложенной модификации, а также показали, что сложность алгоритма осталась прежней.

4. Применение алгоритма к задаче поиска подстрок

В данном разделе мы продемонстрируем, как модификация алгоритма Валианта может быть применена к задаче поиска подстрок.

Пусть мы хотим для входной строки размера найти все подстроки размера , которые принадлежат языку, заданному грамматикой . Тогда мы должны посчитать слои подматриц, размер которых не превышает , где .

Пусть и, следовательно, .

Для всех количество перемножений матриц размера выполняется ровно раз и каждое из них включает перемножение булевых подматриц.

Временная сложность алгоритма для поиска всех подстрок длины равна , где появившийся дополнительный множитель обозначает количество матриц в последнем вычисленном слое, но он, во-первых, мал относительно общей работы алгоритма, во-вторых, не существенен, так как эти матрицы могут быть обработаны параллельно. Алгоритм Валианта, в отличие от модификации, не может так легко быть применен к данной задаче. В нем необходимо будет полностью вычислить, как минимум, две треугольные подматрицы размера (как показано на рис. 5). Это значит, что минимальная сложность, усовершенствовать которую без дополнительных модификаций не удастся, будет составлять .

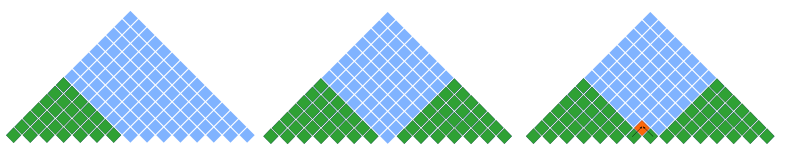


Рис. 5. Количество элементов, которые необходимо вычислить в алгоритме Валианта (2 треугольных подматрицы размера ).

Fig. 5. The number of elements nessesary to compute in Valiant’s algorithm (at least 2 triangle submatrices of size ).

В завершение данного раздела, скажем, что модификация может быть эффективно применена для строк размера , что и было показано в проведённых экспериментах.

5. Эксперименты

В этом разделе мы приводим результаты экспериментов, целью которых является демонстрация практической применимости предложенной модификации алгоритма Валианта к задаче поиска подстрок.

5.1 Постановка экспериментов

Эксперименты проводились на рабочей станции со следующими характеристиками: операционная система —, ЦПУ — Intel i5-8250U, 1600-3400 Mhz, 4 Core(s), 8 Logical Processor(s), оперативная память — 8 GB.

Была выполнена реализация алгоритма Валианта и предложенной модификации на языке программирования С++ []. Также модифицированная версия была адаптирована для задачи поиска подстрок. Для перемножения подматриц использовалась библиотека для высокоэффективной работы с булевыми матрицами M4RI [].

Сначала был проведен сравнительный анализ производительности двух версий алгоритма: исходной и модифицированной.

Затем, для адаптированной под задачу поиска подстрок версии, был проведен подсчет времени, затрачиваемого на одну строку, который позволит сделать вывод о применимости предложенной модификации в областях, работающих с большими объемами данных, например, биоинформатике.

Для экспериментов использовалась КС-грамматика G со стартовым нетерминалом S. Правила данной грамматики представлены на рис.6.

|

|

Рис. 6. Грамматика G

Fig. 6. Grammar G

Представленная грамматика, как и большинство грамматик, используемых в биоинформатике, является сильно неоднозначной и позволяет оценить время работы алгоритма на простом примере*.*

Строки для экспериментов были сгенерированы случайно.

5.2 Результаты

Результаты сравнительного анализа алгоритма Валианта и его модификации приведены представлены в табл. 1, где – длина сгенерированной строки. Для двух реализаций представлено время работы алгоритмов в миллисекундах.

Табл. 1. Результаты сравнительного анализа

Table 1. Evaluation results for comparative analysis

|  |  |  |
| --- | --- | --- |
| *n* | *Valiant’s Algorithm (ms)* | *Modification (ms)* |
| *127* | *39* | *37* |
| *255* | *159* | *153* |
| *511* | *647* | *620* |
| *1023* | *2601* | *2500* |
| *2047* | *10462* | *10080* |
| *4095* | *42004* | *40577* |
| *8191* | *169040* | *164518* |

Результаты работы адаптированной к задаче поиска подстрок модификации представлены в табл.2, где – длина сгенерированной строки, – длина искомых подстрок, – количество подстрок, которое можно найти за этот вызов, – общее время работы адаптированной версии в миллисекундах, – время, затраченное на поиск одной подстроки в миллисекундах.

Табл. 2. Результаты работы алгоритма для задачи поиска подстрок

Table 2. Evaluation results for string-matching problem

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *s* | *n* | *strnum* | *total (ms)* | *per str (ms)* |
| *511* | ***1023*** | *512* | *2500* | *4,9* |
| *511* | ***2047*** | *1536* | *6260* | *4,1* |
| *511* | ***4095*** | *3584* | *13861* | *3,9* |
| *511* | ***8191*** | *7680* | *29331* | *3,8* |
| *1023* | ***2047*** | *1024* | *10080* | *9,8* |
| *1023* | ***4095*** | *3072* | *25320* | *8,2* |
| *1023* | ***8191*** | *7168* | *56683* | *7,9* |
| *2047* | ***4095*** | *2048* | *40577* | *19,8* |
| *2047* | ***8191*** | *6144* | *102722* | *16,7* |

5.3 Анализ результатов

Результаты сравнительного показывают, что предложенная модификация и исходный алгоритм Валианта работают практически одинаково. Небольшой выигрыш модификации будет существенен только на очень длинных строках, на которых последовательная версия работает очень долго.

Результаты работы адаптированной версии показывают, что с ростом размера полной строки, время, затрачиваемое на поиск одной подстроки, непрерывно уменьшается. Данное небольшое преимущество может показать существенное увеличение производительности при работе с большими объемами данных.

Таким образом, поставленные эксперименты демонстрируют практическую применимость предложенной модификации алгоритма Валианта.

6. Заключение

В данной работе был предложен алгоритм

Кроме того, мы можем определить несколько открытых проблем для будущих исследований.

Благодарности

Авторы выражают признательность Кознову Дмитрию Владимировичу за оказанную помощью при написании настоящей статьи. Данная работа выполнена при финансовой поддержке гранта РНФ 18-11-00100 и гранта от JetBrains Research.

Список литературы

1. Okhotin A. 2001. Conjunctive grammars // Journal of Automata, Languages and Combinatorics. 6(4), pp. 519—535.
2. Chomsky N. 1959. On certain formal properties of grammars // Information and control. 2(2), pp. 137—167.
3. Kasami T. 1965. AN EFFICIENT RECOGNITION AND SYNTAXANALYSIS ALGORITHM FOR CONTEXT-FREE LANGUAGES. Technical Report. DTIC Document.
4. Younger D. H. 1967. Recognition and parsing of context-free languages in time n3 // Information and control. 10(2), pp. 189—208.
5. Valiant L. G. 1975. General context-free recognition in less than cubic time // Journal of computer and system sciences. 10(2), pp. 308— 315.
6. Okhotin A. 2013. Conjunctive and Boolean grammars: the true general case of the context-free grammars // Computer Science Review. 9, pp. 27—59.
7. Okhotin A. 2004. Boolean grammars. Information and Computation 194, 1 (2004), pp. 19–48.
8. Okhotin A. 2014. Parsing by matrix multiplication generalized to Boolean grammars. Theoretical Computer Science 516 (2014), pp. 101–120.

**Заголовок статьи на английском языке**

1Y.A. Susanina <[jsusanina@gmail.com](mailto:jsusanina@gmail.com)>

3A.N. Yaveyn <[yaveyn@yandex.ru](mailto:yaveyn@yandex.ru)>

3S.V. Grigorev <[Semen.Grigorev@jetbrains.com](mailto:Semen.Grigorev@jetbrains.com)>

1, 2 Saint Petersburg State University,

7/9, Universitetskaya nab., St. Petersburg, 199034, Russia.

Abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract. Abstract is abstract.

Keywords: matrix operations.

**DOI:** ???

**For citation:** Susanina Y.A., Yaveyn A.N., Grigorev S.V. Title. *Trudy ISP RAN/Proc. ISP RAS*, vol. ?, issue ?, 2019. pp. ?-? (in Russian). DOI: ???

References

1. Okhotin A. 2001. Conjunctive grammars // Journal of Automata, Languages and Combinatorics. 6(4), pp. 519—535.
2. Chomsky N. 1959. On certain formal properties of grammars // Information and control. 2(2), pp. 137—167.
3. Kasami T. 1965. AN EFFICIENT RECOGNITION AND SYNTAXANALYSIS ALGORITHM FOR CONTEXT-FREE LANGUAGES. Technical Report. DTIC Document.
4. Younger D. H. 1967. Recognition and parsing of context-free languages in time n3 // Information and control. 10(2), pp. 189—208.
5. Valiant L. G. 1975. General context-free recognition in less than cubic time // Journal of computer and system sciences. 10(2), pp. 308— 315.
6. Okhotin A. 2013. Conjunctive and Boolean grammars: the true general case of the context-free grammars // Computer Science Review. 9, pp. 27—59.
7. Okhotin A. 2004. Boolean grammars. Information and Computation 194, 1 (2004), pp. 19–48.
8. Okhotin A. 2014. Parsing by matrix multiplication generalized to Boolean grammars. Theoretical Computer Science 516 (2014), pp. 101–120.