ОТЗЫВ

официального оппонента о диссертации Рубинчика Михаила Валентиновича «Вычислительная сложность некоторых задач обработки строк», представленной на соискание учёной степени кандидата физико-математических наук по специальности 01.01.09 — дискретная математика и математическая кибернетика

Актуальность темы. Представленная М. В. Рубинчиком диссертация относится к области алгоритмов на строках. Это активная и бурно развивающаяся область с большим количеством как теоретических результатов, так и практически важных приложений. Алгоритмы на строках используются, например, в информационном поиске, анализе данных, вычислительной биологии, компьютерном зрении. Ежегодно проводится несколько специализированных конференций, посвящённых алгоритмам на строках. Именно благодаря развитию данной области, мы сегодня имеем возможность эффективно анализировать геномные последовательности (в частности, можно скачать специально подготовленный индекс такой последовательности, который позволит осуществить поиск фрагмента в ней очень быстро, без чтения всей последовательности), а также искать нужное нам слово в тексте в режиме реального времени.

Обзор диссертации по главам. В работе строятся новые алгоритмы и структуры данных для вычислительных задач, связанных с палиндромами в строках и повреждениях в тексте.

Первая глава работы является вводной, в ней определяются все необходимые понятия, описывается модель вычисления, а также кратко описываются необходимые для работы существующие алгоритмы и структуры данных.

Во второй главе приводятся алгоритмы и структуры данных для работы с палиндромами в строках. Палиндром — крайне важный объект как области алгоритмов на строках (и теории вычислений вообще), так и области комбинаторики слов. Изучению свойств палиндромов посвящено огромное количество статей. Интересно отметить, что современное название области алгоритмов на строках — «стрингология» — было дано более тридцати лет назад Цви Галилом. По его собственному признанию, его увлечение данной

областью началось с чтения результата Анатолия Олесьевича Слисенко, доказавшего, что существует алгоритм, распознающий палиндромы в реальном времени.

В работе рассматриваются такие известные в данной области задачи, как разбиение строк на палиндромы, поиск числа различных палиндромов в строке и поиск числа палиндромно-насыщенных (имеющих наибольшее число различных палиндромов) строк. По всем этим задачам приводятся новые решения, которые либо асимптотически лучше всех предшественников, либо имеют лучшую константу в практических реализациях. Главным результатом главы является структура данных «овердрево» (eertree), которая позволяет эффективно хранить данные о подпалиндромах строки (или нескольких строк). Именно с её помощью автором получены улучшения в перечисленных выше задачах. Появившийся ранее алгоритм Манакера имеет возможность хранить информацию о том, какая подстрока данной строки является палиндромом, а какая нет, однако не хранит информацию о том, какие палиндромы равны, а какие нет. Поэтому во многих задач о палиндромах для решения использовались одновременно алгоритм Манакера и суффиксные структуры данных. Овердрево является существенно более простой, чем суффиксные структуры, однако для палиндромов в каком-то смысле выполняет одновременно роль алгоритма Манакера и любой суффиксной структуры. В работе приводятся различные модификации овердрева, позволяющие получить овердрево с откатами и персистентное овердрево, приводится ряд задач, позволяющих понять особенности устройства овердрева. Кроме того, с помощью овердрева улучшены решения нескольких известных задач: нахождение всех палиндромов строки, разбиение строки на палиндромы, перечисление всех палиндромно-насыщенных строк.

Третья глава посвящена работе с повреждёнными строками. Повреждённые строки — строки, в которых некоторые символы являются символами алфавита, а некоторые «повреждены». Для каждого повреждённого символа известен набор символов алфавита, в которые его можно «восстановить». Известным частным случаем повреждённых строк являются частичные строки. В них все повреждённые символы повреждены «полностью», т. е. нет информации о том, в какой именно символ можно восстановить каждый из них. Таким образом, каждый повреждённый символ можно восстановить в любой. Известны задачи о поиске шаблона в тексте для повреждённых строк,

также известно, что существуют решения этих задач с помощью булевых конволюций. Автором ставится родственная задача о восстановлении текста и шаблона по некоторому критерию. В работе рассматривается два критерия: максимизация числа вхождений шаблона в текст и минимизация суммарного расстояния Хэмминга от шаблона до всех подстрок текста соответствующей длины. Автор приводит доказательства NP-трудности обеих задач в общем случае (для первой даже для бинарного алфавита), однако оказывается, что для многих интересных частных случаев задачи имеют полиномиальные решения, что и показано в работе. Для первой задачи приводятся решения в случаях неповреждённого шаблона и неповреждённого текста. Во второй задаче помимо случаев неповреждённых шаблона или текста, разобраны случай бинарного алфавита, случай частичных строк с цикличным текстом. Также сформулирована гипотеза об NP-трудности второй задачи для случая частичных строк и обычного (не циклического) текста.

Замечания к работе. Основным недостатком по содержанию работы, на мой взгляд, является доказательство NP-полноты задачи минимизации расстояния Хэмминга (задача 2 в третьей главе). Доказательство NP-полноты приведено с помощью сведения задачи MAX-SAT в случае, когда текст является циклической строкой. В большинстве приложений, как правило, длина текста значительно превосходит длину шаблона, в сведении же шаблон несколько раз может наматываться на циклический текст, что вызывает ощущение искусственности рассматриваемой задачи и не даёт понимания о сложности задачи на естественных входах.

Что касается оформления, в работе имеется незначительное число опечаток:

1. Страница 9, строка 13: «равный её равный префиксу»
2. Страница 42, строка 3: вместо «$\log \delta$» должно быть «$\log \Sigma$»
3. Страница 65, строка 1 снизу: вместо «$v_j = 1w_j10^{k-1}$» должно быть «$v_j = 1w_j10^{k+1}$»

Заключение. Работа демонстрирует глубокое понимание автором области алгоритмов на строках и теории NP-полноты и высокое владение матема-

тическим аппаратом. Диссертация чётко структурирована, все утверждения обоснованы и корректность доказываемых теорем не вызывает сомнений. Все результаты диссертации получены автором впервые и опубликованы, в том числе, в ведущих рецензируемых научных журналах, определённых ВАК. Результаты неоднократно докладывались на российских и международных конференциях. Автореферат соответствует содержанию диссертации.

Проведённое рассмотрение диссертации показывает, что данная работа Рубинчика М. В. отвечает всем критериям раздела II Положения о присуждении ученых степеней, установленным для кандидатских диссертаций по специальности 01.01.09 — дискретная математика и математическая кибернетика, а ее автор заслуживает присуждения искомой ученой степени.

29.04.2016

Официальный оппонент:
к. ф.-м. н., научный сотрудник лаборатории алгоритмических методов
ФГБУН «Санкт-Петербургское отделение Математического института
им. В. А. Стеклова РАН», 191023, Санкт-Петербург, наб.р. Фонтанки, д. 27
kulikov@logic.pdmi.ras.ru,
+7 (812) 312-40-58 А. С. Куликов

Подпись официального оппонента заверяю:
директор Санкт-Петербургского отделения
Математического института им. В. А. Стеклова РАН,
член–корреспондент РАН С. В. Кисляков