ПРЕПРИНТ ПОМИ — 22/2006

РАЗРЕЖЕННАЯ ПЕРИОДИЧНОСТЬ*

Юхани Карьюмяки
Университет Турку
http://users.utu.fi/karhumak/

Юрий Лифшиц
Петербургское отделение Математического института
им. В.А. Стеклова Российской академии наук
http://logic.pdmi.ras.ru/~yura/

16 декабря 2006 г.

Аннотация

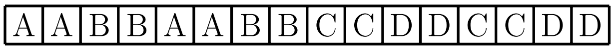
В этой работе мы обобщаем классическое понятие периодичности. Слово называется разреженно периодическим, если оно может быть разделено на несколько копий частично определенного слова. Например, слово XXYY является разреженно периодическим, так как оно состоит из двух копий частично определенного слова X_Y. Мы изучим это понятие, обсудим его свойства и представим алгоритм поиска минимальных разреженных периодов. В процессе исследований нам удалось ответить на два открытых вопроса, поставленных соответственно Теро Харью и Нейлом Слоаном.

*Работа первого автора поддержана грантом N 8206039 Финской Академии, второй автор благодарит за поддержку программы INTAS 04-77-7173 и НШ-8464.2006.1.

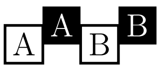
1 Введение

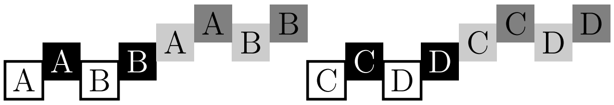
Основное понятие. Слово S называется *чисто периодическим*, если $S = W^k = W \dots W$. Другими словами, чистая периодичность соответствует делимости $p|n$ и равенству $s_i = s_{i+p}$ для всех $1 \leq i < i+p \leq n$. В этой работе мы будем иметь дело только с чистой периодичностью.

Два слова на иллюстрации внизу не являются чисто периодическими. Однако, они обладают некоторой структурной закономерностью. В этой работе мы вводим новое обобщенное понятие периодичности, в соответствии с которым эти строки становятся периодическими.



Неформально, слово T называется *разреженно периодическим*, если существует такое частично определенное слово S, что T может быть разделено на несколько параллельных копий S. Мы представим формальное определение со всеми подробностями в Разделе 2. На следующей иллюстрации показаны разбиения, доказывающие разреженную периодичность исходных примеров.





Рассматриваемые вопросы. Эта работа посвящена исследованию основных свойств разреженной периодичности и сопоставлению их с классическим определением. Нас интересуют следующие вопросы: (1) Как перечислить все разреженные периоды? (2) Каждое ли слово имеет единственный *примитивный* разреженный период? (3) Как найти все разреженные периоды минимального размера? (4) Сколько разреженных периодов может иметь слово длины n? (5) Каково отношение между примитивным классическим периодом и примитивным разреженным периодом?

У нас есть несколько причин изучать разреженные периоды. Во-первых, это естественное обобщение классического понятия, так как любой классический период является также и разреженным. Обратное не обязательно верно. Во-вторых, в случае, когда разреженный период относительно мал, мы можем полностью описать длинную строку сообщив только ее длину и разреженный период. Таким образом, мы получаем новый формализм для описания класса слов с малой колмогоровской сложностью. Разреженная периодичность может быть полезна для построения новых алгоритмов архивирования (особенно для обобщения run-length encoding). Заметим здесь, что отношение между “размером” разреженного периода и длиной классического периода может быть сколь угодно мало. В-третьих, понятие разреженной периодичности дает геометрический взгляд на структуру текстов. Далее, в Разделе 5 мы формулируем гипотезу, утверждающую что разреженная периодичность не выразима при помощи уравнений в словах. Еще одним стимулом для изучения разреженной периодичности является надежда обнаружить новые закономерности в реальных данных. Представляется интересным провести экспериментальные исследования в этом направлении.

Покажем один естественный источник разреженной периодичности. Рассмотрим многомерный параллелепипед $n_1 \times \cdots \times n_d$. Пусть каждая его клетка покрашена в некоторый цвет. Отсортируем все клетки в алфавитном порядке их координат (от $(0, \ldots, 0)$ до $(n_1 - 1, \ldots, n_d - 1))$ и выпишем все их цвета в единую последовательность. Предположим теперь, что исходный параллелепипед был замощен меньшим $m_1 \times \cdots \times m_d$, причем $m_1 | n_1, \ldots, m_d | n_d$ и каждая копия маленького параллелепипеда была раскрашена одинаково. Тогда последовательность цветов большого параллелепипеда будет обладать разреженной периодичностью. Мы вернемся к этому примеру после введения ряда полезных терминов. Другими источниками разреженной периодичности могут стать XML-файлы и автоматически порожденные тексты.

Результаты. Первым результатом работы является уже само понятие разреженной периодичности. Оно выглядит очень просто и естественно, но, насколько нам известно, никогда не было явно сформулировано прежде.

Далее, мы ввели частичный порядок на разреженных периодах и обнаружили, что, в отличие от классического случая, у текста может оказаться несколько примитивных разреженных периодов. Мы представим наименьший из известных нам примеров (24 буквы), обладающих этим свойством, в Разделе 3. Этот пример позволил опровергнуть гипотезу Теро Харью об *общем подразбиении*.

Тем не менее, в Подразделе 4.2 мы докажем что каждый примитивный разреженный период слова T также является разреженным периодом классического примитивного периода T. Из этого свойства следует, что разреженная периодичность живет “внутри” классического примитивного периода. Доказательство основано на расширенном алгоритме Евклида.

В Подразделе 4.1 мы представим классификацию всех возможных разреженных периодов. Наиболее трудным техническим шагом является доказательство полноты нашей конструкции, то есть необходимость показать, что каждый разреженный период в самом деле включен в нашу классификацию. Мы получили рекурсивную формулу для вычисления функции $L(n)$, которая дает максимальное количество разреженных периодов для слов длины n. Значение $L(n)$ может оказаться даже больше чем длина исходного текста. Например, слово длиной 36 символов может иметь до $L(36) = 52$ разреженных периодов. Попутно мы доказали что последовательности A067824 и A107736 из Энциклопедии целочисленных последовательностей [16] совпадают.

Наконец, в Подразделе 4.3 мы представим алгоритм для нахождения (всех) разреженных периодов минимального размера для данного текста. Трудоемкость алгоритма составляет $n^{1+o(1)}$ шагов.

Английский перевод этой работы размещен на домашней интернет-странице второго автора http://logic.pdmi.ras.ru/~yura/.

1.1 Обзор близких результатов

Одним из наиболее важных результатов для классической периодичности является теорема Файна и Вильфа [7]. Они изучали необходимые и достаточные условия, при которых из p-периодичности и q-периодичности следует НОД(p, q)-периодичность (НОД = наибольший общий делитель). Мы предприняли попытку найти аналогичное свойство для разреженных периодов. К нашему удивлению, теорема Файна и Вильфа не обобщается на разреженную периодичность. Даже два *чистых* разреженных периода одного текста могут не иметь никакого общего подпериода.

Наше обобщение понятия периодичности является далеко не первым. Недавно Симпсон и Тайдеман обобщили теорему Файна и Вильфа на многомерные периодичности [15]. Они считают, что некоторый вектор $\bar{v}$ является периодом многомерной раскрашенной клетчатой фигуры, если для любая клетка x покрашена в тот же цвет, что и $x + \bar{v}$. Другое обобщение теоремы Файна и Вильфа можно найти в статье [5].

Берстель и Боассон [10], и вслед за ними Шур и Коновалова [14], Бланше-Садри [2, 3, 4] подробно изучили частичные слова и ввели понятие периодичности для них. Однако, в этой постановке не допускается перекрытие периодов (копии периода должны идти одна вслед за другой). Таким образом, в этой модели только частичные слова могут иметь разреженные периоды (т.е. периоды с пропусками букв).

В работе [1] Апостолико, Фарач и Илиопулос ввели понятие накрытия (cover). Слово C называется накрытием для слова T, если каждая буква из T находится внутри некоторого вхождения C в T. Таким образом, эта модель отличается от нашего понятия в двух аспектах: накрытия связны (никаких пропусков) и накрытия могут перекрываться.

Катона и Шаш в своей работе [12] предложили вариант разреженной периодичности на плоскости. К сожалению, они рассмотрели только периоды (шаблоны) из двух клеток.

Наконец, понятие периодичности для непрерывных функций было обобщено под именем *почти периодических функций* [6].

2 Понятие разреженной периодичности

Частично определенное слово — это слово в алфавите $\Sigma \cup \{\diamond\}$, где $\diamond$ — это специальная прозрачная (или неопределенная) буква. Другими словами, частично определенное слово представляет собой последовательность обычных слов (блоков), разделенных пропусками фиксированной (но необязательно одинаковой) длины.

Частично определенное слово S называется *разреженным периодом* (обычного) слова T, если мы можем разделить T на несколько параллельно сдвинутых копий S, которые будут удовлетворять следующим условиям:

- Все определенные (видимые) буквы S-копий совпадают с соответствующими буквами в тексте;

- Каждая буква текста покрыта *в точности одной* определенной (видимой) буквой из S-копий.

Представим, что у нас есть несколько копий частично определенного слова, напечатанных на прозрачной бумаге. Тогда это слово будет разреженным периодом некоторого текста, если мы можем сложить наши копии в стопку так, чтобы образовалась одна связная строка без перекрытий видимых букв.

Пример: X◇Y является разреженным периодом для XXYY.

3 Примитивный разреженный период не единственен

Частичный порядок на частично определенных словах. Будем говорить, что одно частичное слово S *меньше* чем другое частичное слово Q, если Q может быть разделено на несколько параллельно сдвинутых копий S так, чтобы были выполнены следующие свойства:

- Все определенные (видимые) буквы S-копий совпадают с соответствующими буквами Q

- Каждая буква Q покрыта *в точности одной* определенной буквой из S-копий.

В частности, любой разреженный период меньше чем сам текст. На самом деле, это определение просто расширяет разреженную периодичность на частично определенные слова.

Пример:

A	A	B	B	A	A	B	B	C	C	D	D	C	C	D	D

меньше чем

A	A	B	B	A	A	B	B	C	C	D	D	C	C	D	D

В классической периодичности для каждого текста существует единственный примитивный чистый период. Здесь под словом “примитивный” мы имеем в виду, что любой другой период является всего лишь последовательным повторением нескольких копий примитивного периода. Для обнаружения разреженной периодичности было бы крайне важно иметь такое же свойство. Сформулируем его ниже.

“Вопрос о примитивном разреженным периоде”: верно ли, что для каждого обычного слова существует единственный примитивный разреженный период (то есть он меньше любого другого)? Переформулировка: верно ли, что любые два разреженных периода одного слова имеют общий разреженный “подпериод”?

Удивительно, но эта столь естественно выглядящая гипотеза неверна. Здесь мы представляем наименьший из известных нам (24 буквы) контрпримеров. Следующие разреженные периоды

A	A	A	A	A	A	A	A	B	A	A	B	B	A	A	B	A	A	A	A	A	A	A	A

и

A	A	A	A	A	A	A	A	B	A	A	B	B	A	A	B	A	A	A	A	A	A	A	A

не сравнимы между собой. Дадим лишь краткий набросок объяснения, почему у них не может быть общего подпериода. Предположим, что он все же есть. Тогда это частично определенное слово имеет только одну букву B. Так как расстояние между первой и второй буквами B в тексте равно трем, сдвиг между первой и второй копией нашего подпериода тоже равен трем. Таким образом, это частично определенное слово начинается с трех подряд идущих A. Но с другой стороны, сдвиг третьей копии относительно второй равен единице. И следовательно, стартовые блоки второй и третьей копии перекрываются.

Как получить больше двух попарно несравнимых разреженных периодов? Мы можем использовать для этой цели "рекурсивную конструкцию". Пусть T — это текст вышеприведенного контрпримера, а тексты $T_{A_1B_1}$ и $T_{A_2B_2}$ получены из T просто использованием других букв. Построим текст T_2, заменяя каждую букву A в T на $T_{A_1B_1}$ и каждую B в T на $T_{A_2B_2}$. Теперь мы можем описать четыре несравнимых периода для T_2. Сам текст T_2 имеет длину 24^2. Сгруппируем его буквы на 24 блока по 24 буквы. Оставим лишь 12 блоков, используя как шаблон одно из двух вышеприведенных частично определенных слов. Теперь внутри каждого блока тоже оставим лишь 12 символов, используя один из двух разреженных периодов исходного контрпримера (одинаково в каждом блоке).

Мы можем повторять эту конструкцию несколько раз. Тогда текст T_k будет иметь длину 24^k и у него будет 2^k попарно несравнимых периодов. Асимптотически, эта конструкция дает нижнюю оценку $n^{\frac{\log 2}{\log 24}} > \sqrt[5]{n}$ на количество примитивных разреженных периодов.

В 2003 году в своем курсе лекций [8] Теро Харью поставил следующий вопрос. Предположим, что некоторая раскрашенная клетчатая фигура имеет разбиение на одинаковые копии одного шаблона, и может быть разбита на копии другого шаблона. Верно ли, что всегда существует такой третий шаблон, что на его копии можно разбить и первый, и второй? Этот вопрос был мотивирован изучением defect effect в комбинаторике на словах [9]. Наш пример дает отрицательный ответ на вопрос Харью даже для одномерных (не обязательно связных!) фигур.

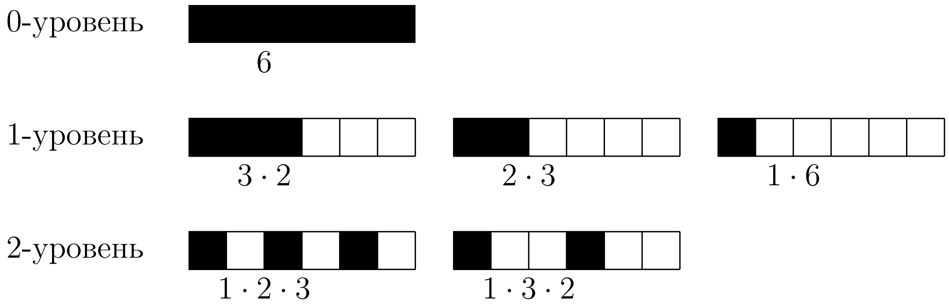
4 Свойства разреженной периодичности

Начнем наше исследование с классификации всех возможных разреженных периодов унарного (используется только одна буква) слова длины n. Так мы получим точную верхнюю оценку на количество всех разреженных периодов для текста длины n. Далее мы переформулируем разреженную периодичность в алгебраических терминах (аналогично правилу $s_i = s_{i+p}$ для классической периодичности). Эта переформулировка поможет нам доказать, что каждый примитивный разреженный период меньше классического примитивного периода, и позволит создать алгоритм по поиску всех разреженных периодов минимального размера.

4.1 Количество разреженных периодов

Теорема 1. *Существует биективное соответствие между разреженными периодами унарного слова длины n и упорядоченными разложениями $n = n_1 \cdot \ldots \cdot n_k$, где $n_2, \ldots, n_k \geq 2$.*

Доказательство. Мы начнем с описания множества разреженных периодов (иерархическая конструкция) унарного слова длины n. После этого мы докажем полноту этого множества, то есть что любой разреженный период в него входит.



Мы разделяем множество всех разреженных периодов на уровни. Каждому периоду в нашей системе будет поставлен в соответствие специальный *код*. Весь текст будет единственным периодом 0-уровня и ему будет соответствовать код n. Для каждого разложения $n = n_1 \cdot n_2$, где $n_2 \geq 2$, блок из первых n_1 букв будет периодом 1-уровня с кодом $n_1 \cdot n_2$. На самом деле, 0-уровень и 1-уровень вместе составляют множество всех классических (то есть связных) периодов.

Объясним теперь, как из периода P из k-уровня и с кодом $n_1 \cdot \ldots \cdot n_k$ построить новый разреженный период Q из $k+2$-уровня. Нужно взять разложение $n_1 = m_1 \cdot m_2 \cdot m_3$, где $m_2, m_3 \geq 2$. В течение всего процесса построения первое число в коде равно длине непрерывных блоков в периоде (в рамках наших построений все блоки в периоде имеют равную длину). Для построения нового разреженного периода Q мы возьмем каждый блок P, разделим его на m_3 групп по $m_1 \cdot m_2$ символов, и в каждой такой группе оставим только m_1. Заметим здесь, что P может быть покрыто m_2 копиями Q (со сдвигами $0, m_1, \ldots, m_1 \cdot (m_2 - 1)$). Таким образом, Q также является разреженным периодом. Мы присвоим ему код $m_1 \cdot m_2 \cdot m_3 \cdot n_2 \cdot \ldots \cdot n_k$. Подчеркнем, что наше построение сохраняет неравенство $n_2, \ldots, n_k \geq 2$ для всех кодов.

Докажем теперь, что каждый разреженный период включен в наше построение. Возьмем произвольный разреженный период P. Все его блоки должны иметь одинаковую длину s, и длина каждого пропуска должна быть кратна s. В самом деле, в замощении текста вторая копия P сдвинута на длину первого блока s_1. Таким образом, чтобы избежать перекрытия все другие блоки по длине не превосходят s_1. Предположим теперь, что какой-то блок b оказался первым из тех, чья длина все же *строго* меньше чем s_1. Тогда длина пропуска между блоком b из первой копии P и b из второй копии P строго меньше s_1. Следовательно, этот пропуск не может быть заполнен. Каждый пропуск в P заполнен несколькими связными блоками из других копий. Следовательно, все пропуски кратны s.

Нам потребуется определить *простую степень* для каждого разреженного периода P. Пусть s — это размер блока, а $g \cdot s$ — длина первого пропуска. Простой степенью P будет объединение $g+1$ копии P со сдвигами $0, s, 2s, \ldots, g \cdot s$.

Утверждение: Простая степень произвольного разреженного периода P также является разреженным периодом.

Доказательство: возьмем разбиение текста на P-копии. Пусть s это размер блока, а $g \cdot s$ — длина первого пропуска в P. Разделим все копии P на группы по $g+1$ экземпляров следующим образом. Будем рассматривать все копии слева направо. Пропуск между первыми двумя блоками первой копии P может быть заполнен *только первыми* блоками других P-копий. Первый экземпляр P и участники заполнения первого пропуска составят нашу первую группу. Предположим теперь, что у нас уже создано несколько групп. Рассмотрим первую еще не использованную копию P. Посмотрим на ее первый пропуск. Все P-копии, включенные в уже созданные группы, идут чересчур длинными (не менее $s \cdot (g+1)$) непрерывными участками. Таким образом, рассматриваемый пропуск тоже заполнен новыми, еще не использованными копиями. Так как мы обрабатываем копии строго слева направо, для заполнения нашего пропуска используются только первые блоки других копий. Следовательно, мы можем объединить их в очередную группу из $g+1$ копии, идущей непрерывно одна за другой. Отметим, что каждая группа в точности является простой степенью P. Таким образом, мы доказали, что исходный текст имеет разбиение на копии простой степени. Утверждение доказано.

Если существует хоть один разреженный период P вне нашей конструкции, то будет и такой разреженный период P', что (1) сам он не включен в наше построение и (2) его простая степень Q входит в нашу иерархию. Выведем противоречие из этих посылок. В самом деле, пусть $n_1 \cdot \ldots \cdot n_k$ — код Q, длина блоков P' равна s, а $s \cdot g$ равно длине первого пропуска. Тогда размер блока n_1 в разреженном периоде Q кратен $s \cdot (g+1)$. Пусть $m_1 = s, m_2 = g+1, m_3 = n_1/(s \cdot (g+1))$. Остается заметить, что P' на самом деле является разреженным периодом нашей иерархии с кодом $m_1 \cdot m_2 \cdot m_3 \cdot n_2 \cdot \ \ldots \ \cdot n_k$.

Теорема доказана. □

Следствие 1. Пусть $L(n)$ — это количество разреженных периодов унарного слова длины n. Только что доказанная теорема утверждает, что $L(n)$ равно количеству разложений $n = n_1 \cdot \cdots \cdot n_k$, где $n_2, \ldots, n_k \geq 2$. Группируя все разложения по крайне правому сомножителю, мы получим следующую рекурсивную формулу:

$$L(1) = 1; \qquad\qquad L(n) = 1 + \sum_{d|n, d \neq n} L(d).$$

Замечание. Энциклопедия целочисленных последовательностей [16], поддерживаемая Нейлом Слоаном, содержит две интересные для нас последовательности. Последовательность A067824 определяется в точности формулой из Следствия 1. Последовательность A107736 выражает количество многочленов p с коэффициентами из $\{0,1\}$, которые делят $x^n - 1$, и таких, что коэффициенты частного при делении $(x^n - 1)/((x-1)p(x))$ также принадлежат $\{0,1\}$. Но это число в точности равно количеству разреженных периодов унарного слова длины n. В самом деле, умножая p на многочлен с коэффициентами из $\{0,1\}$, мы проводим разбиение n "единиц" на несколько параллельно сдвинутых копий p без перекрытий. В Энциклопедии Слоана отмечено, что последовательности A067824 и A107736 совпадают по крайней мере на первых 300 значениях, но никакого доказательства полного совпадения до сих пор известно не было. Теорема 1 и Следствие 1 дают такое доказательство. Действительно, количество многочленов равно количеству разреженных периодов, количество разреженных периодов равно числу разложений (Теорема 1), а число разложений удовлетворяет нужной рекурсивной формуле (Следствие 1).

Функция $L(n)$ и число разложений также появлялись в книге Дональда Кнута "Искусство программирования" [13]. Тем не менее, никакой замкнутой формулы для вычисления $L(n)$ пока не известно. Простыми вычислениями мы можем показать, что для всяких простых чисел p и q верны следующие равенства:

- $L(p) = 2$,
- $L(p^k) = 2^k$,
- $L(pq) = 6$,
- $L(p^2q) = 1 + L(1) + L(p) + L(q) + L(pq) + L(p^2) = 16$,
- $L(p^2q^2) = 52$, в частности $L(36) = 52$.

Как мы видим, слово длины 36 может иметь до 52 разреженных периодов. Следовательно, найдутся два разных периода с одинаковым "размахом" (расстоянием между первой и последней видимой буквой).

Теперь мы можем сказать больше и о примере во введении. Если d-мерный параллелепипед разбит на копии меньшего параллелепипеда, то последовательность цветов будет обладать разреженным периодом $2d$-уровня. Размер этого периода (число видимых букв) будет равно объему меньшего параллелепипеда.

4.2 Соотношение между разреженными и классическими периодами

Определение. Будем говорить,что слово $T = T_0 \dots T_{n-1}$ имеет *чистый ограниченный период* (a, b) (или период a с ограничением b), если $b|n$, $a|b$, $a < b$ и выполнено следующее утверждение:

$$\left\lfloor \frac{i}{b} \right\rfloor = \left\lfloor \frac{i+a}{b} \right\rfloor \Rightarrow T_i = T_{i+a}.$$

Другими словами, если мы разделим весь текст на блоки длины b, то каждый из блоков будет иметь чистый период a. Классический период p можно описать как ограниченный период (p, n).

A	A	B	B	A	A	B	B	C	C	D	D	C	C	D	D

Этот текст имеет ограниченные периоды $(1, 2)$ и $(4, 8)$.

Лемма 1. *У слова T есть разреженный период P с кодом $n_1 \cdot \ldots \cdot n_{2k}$ тогда и только тогда, когда текст обладает ограниченными периодами*

$$(n_1, n_1 n_2), \ldots, (\prod_{i=1}^{2k-1} n_i, \prod_{i=1}^{2k} n_i).$$

У слова T есть разреженный период P с кодом $n_1 \cdot \ldots \cdot n_{2k+1}$ тогда и только тогда, когда текст обладает ограниченными периодами

$$(n_1, n_1 n_2), \ldots, (\prod_{i=1}^{2k-1} n_i, \prod_{i=1}^{2k} n_i).$$

Доказательство. Шаг 1: от разреженного периода к ограниченным периодичностям. Будем использовать индукцию по уровню разреженного периода. Рассмотрим соответствующее разбиение текста. Напомним, что все копии P могут быть разделены на группы по n_2 экземпляров, со сдвигами внутри группы $0, n_1, \ldots, n_1(n_2 - 1)$. Таким образом, если мы разделим весь текст на блоки по $n_1 n_2$ букв, каждый блок будет покрыт P-копиями из одной группы. Следовательно, внутри каждый блок будет n_1-периодическим. Таким образом, мы доказали $(n_1, n_1 n_2)$-ограниченную периодичность. Все остальные ограниченные периодичности следуют из индукционного предположения для простой степени P.

Шаг 2: от ограниченных периодичностей к разреженному периоду. Рассмотрим *верхнюю* ограниченную периодичность $(\prod_{i=1}^{2k-1} n_i, \prod_{i=1}^{2k} n_i)$. Возьмем частичное слово Q, соответствующее коду $(\prod_{i=1}^{2k-1} n_i) \cdot n_{2k}$ для четного варианта леммы, и соответствующее коду $(\prod_{i=1}^{2k-1} n_i) \cdot n_{2k} \cdot n_{2k+1}$ для нечетного варианта. Напрямую из верхней ограниченной периодичности следует, что каждая копия Q на одинаковых позициях имеет одинаковые буквы. Следовательно, Q является разреженным периодом текста. Продолжая рассуждение, с помощью второй ограниченной периодичности можно построить следующий разреженный период R внутри Q. В конце концов, из последней ограниченной периодичности мы сможем вывести существование разреженного периода с требуемым кодом. □

Лемма 2. *Пусть у текста T есть классический период p и ограниченный период (a, b), тогда либо $b|p$ или у текста есть также классический период НОД(a, p).*

Доказательство. Возьмем произвольную букву текста T_i. Мы будем стараться сделать несколько прыжков величиной $\pm p$ и $+a$, чтобы в итоге достичь позиции $i +$ НОД(a, p). Нам хочется все время находиться на позициях с одинаковыми буквами. Поэтому мы запретим себе делать прыжок $+a$ из последнего a-блока в каждом b-блоке. Согласно свойствам расширенного алгоритма Евклида, существуют такие натуральные числа k и l, что НОД$(a, p) = ka - lp$. Мы будем использовать следующую *жадную* стратегию. Если мы имеем право сделать прыжок $+a$, мы его делаем. В противном случае, попытаемся сделать несколько прыжков $\pm p$. После того, как мы сделаем в точности k прыжков $+a$, нам останется лишь провести все недостающие передвижения $\pm p$, чтобы очутиться на желаемой позиции. Единственной преградой на пути жадной стратегии может стать ситуация, когда для некоторого $j < p$ для каждой из позиций $j, j+p, \dots j + (\frac{n}{p} - 1)p$ нам запрещен прыжок $+a$. То есть случай, когда все эти позиции попали в последние a-блоки в своих b-блоках. Предположим теперь, что p не делится на b. Тогда $u =$ НОД$(p, b) \leq \frac{b}{2}$. Так как $p|n$ и $b|n$, среди чисел $j \bmod b, \dots, [j + (\frac{n}{p} - 1)p] \bmod b$ встречаются все такие остатки h по модулю b, что $h \equiv j (\mathrm{mod}\ u)$. Следовательно, один из них не больше u, что в свою очередь не превосходит $b/2$. Но это значит, что соответствующая позиция никак не могла попасть в последний a-блок своего b-блока. □

Теорема 2. *Любой примитивный разреженный период Q слова T является также разреженным периодом для классического примитивного периода T.*

Доказательство. Воспользуемся индукцией по длине текста. Для однобуквенных слов теорема верна. Пусть p — длина классического периода, а (a, b) — верхняя ограниченная периодичность (полученная из кода Q по Лемме 1). Если $p = n$, теорема верна, поэтому будем считать, что $p < n$. Согласно иерархической конструкции, все видимые буквы частично определенного слова Q находятся в первых a-блоках своих b-блоков. Применим Лемму 2. Если $b|p$, то мы можем получить новый разреженный период Q', оставив от Q только ту часть, которая входит в первые p символов текста. Согласно p-периодичности Q может быть разбито на несколько копий Q' со сдвигами $0, p, 2p, \dots, (\frac{n}{p} - 1)p$. Мы получили противоречие с примитивностью Q.

Таким образом, верно второе утверждение из Леммы 2: текст обладает классическим НОД(a, p)-периодом. Но по условию теоремы, p является примитивным классическим периодом. Следовательно, $p|a$. А раз текст является p-периодическим, то он также и a-периодичен и b-периодичен. На этот раз мы можем оставить от Q только буквы, входящие в первый a-блок текста. Этим мы либо получим меньший разреженный период Q' (противоречие с примитивностью Q), или $b = n$ и все видимые буквы Q расположены в первом a-блоке текста. В последнем случае, и классический период p, и разреженный Q являются периодами первого a-блока. Так как $a \leq b/2 \leq n/2$, мы можем применить индукционное предположение. □

Следствие 2. Для каждого разреженного периода Q и классического периода p найдется общий разреженный подпериод.

Доказательство. Возьмем примитивный разреженный период Q', который меньше (то есть разбивает) Q. Рассмотрим классический примитивный период p'. Из единственности чистого классического примитивного периода следует $p'|p$. По Теореме 2 разреженный период Q' является также разреженным периодом для первого p'-блока текста. Таким образом, Q' является искомым разреженным подпериодом для Q и p. □

Замечание. Используя технику из Леммы 1, Леммы 2 и Теоремы 2, можно показать, что примитивный разреженный период единственен в случае $n = 2^k$. Приведем здесь лишь набросок доказательства. От противного: возьмем два несравнимых примитивных разреженных периода. Рассмотрим их верхние ограниченные периодичности. Применим рассуждения, подобные Лемме 2. Либо один из этих периодов не примитивен, либо эти верхние ограниченные периодичности совпадают. Двигаясь дальше вниз, мы либо докажем их равенство, либо получим противоречие с примитивностью.

4.3 Алгоритм поиска разреженных периодов минимального размера

Определим размер разреженного периода как число его видимых букв. Как найти все разреженные периоды минимального размера? Будем пользоваться Леммой 1 для ответа на этот вопрос. Мы знаем, что каждому разреженному периоду соответствует цепочка *вложенных* ограниченных периодичностей $(a_1, b_1), \ldots, (a_k, b_k)$. Под "вложенностью" мы понимаем делимость $b_i|a_{i+1}$ для каждого i. Заметим, что размер такого разреженного периода равен произведению $n \prod_{i=1}^{k} \frac{a_i}{b_i}$.

Алгоритм поиска разреженных периодов минимального размера.

1. Для каждых a, b, таких что $a|b$ и $b|n$, проверить, обладает ли текст (a, b)-ограниченной периодичностью.
2. Построить решетку (ациклический граф) из найденных ограниченных периодичностей. Пары (a, b) будут вершинами, и мы будем проводить ребро от (a, b) к (c, d), если $b|c$.
3. В каждую вершину (a, b) записать значение a/b.
4. Найти все пути в графе с наименьшим произведением значений в вершинах.

Анализ трудоемкости. Первый шаг требует $O(d^2(n)n)$ времени при переборном подходе. Здесь $d(n)$ обозначает количество делителей n. Четвертый шаг лишь незначительно отличается от поиска кратчайших путей в ациклических графах. Сначала нужно совершить один проход по всему графу сверху (маленький период, маленькое ограничение) вниз (большой период, большое ограничение). При проходе мы будем стирать все ребра, которые не участвуют в получении минимальных произведений по путям. В итоге мы получим несколько *стоков* (вершин без исходящих ребер), мы оставим только тот (или те) из них, которые дают минимальное произведение по пути, и удалим все остальные стоки и ребра ведущие к ним. После этого нужно пройти весь граф обратно снизу вверх и стереть все ребра, не участвующие в минимальных путях. В оставшемся графе каждый путь из некоторого истока (вершины без входящих ребер) к некоторому стоку соответствует разреженному периоду минимального размера. Таким образом, мы получили сжатое представление всех разреженных периодов минимального размера. Четвертый шаг алгоритма требует лишь линейное время относительно размера графа. Итоговая трудоемкость алгоритма составляет $O(d^2(n)n + d^2(n)) = n^{1+o(1)}$.

5 Направления будущих исследований

Мы можем предложить множество естественных, важных и, возможно, не столь сложных вопросов для дальнейшего исследования разреженных периодов. Приведем список этих задач.

1. Изучить *не обязательно чистую* разреженную периодичность.
2. Найти точную асимптотическую верхнюю границу $L(n)$.
3. Построить более эффективный (линейный?) алгоритм для нахождения разреженного периода минимального размера.
4. Вычислить, как часто случайное слово (в разных моделях) обладает собственным разреженным периодом. Сравнить полученные результаты с классическим случаем.
5. Проверить, можно ли выразить свойство *"слово имеет разреженный квадратный корень"* через уравнения в словах. Мы отсылаем читателя к работе [11] для знакомства с исследованиями по выразительности уравнений в словах.
6. Проверить, верно ли, что все примитивные периоды имеют одинаковое число видимых букв.
7. Найти другие естественные источники разреженной периодичности.
8. Определить и изучить разреженную периодичность на деревьях и других объектах.
9. Определить и изучить приближенную разреженную периодичность (допускающую небольшие ошибки).

10. Определить и изучить разбиения на копии двух (или более) частичных слов. Этот тип разбиений может оказаться еще более подходящим для применений в архивировании.

На наш вкус, первый вопрос является наиболее важным, так как такое расширение понятия разреженной периодичности позволит описать намного больше структурных свойств текстов. Кроме решения вышеприведенных вопросов необходимо найти близкие результаты и техники в комбинаторике. Такое естественное понятие, как разреженная периодичность, просто обязано иметь предшественников в дискретной математике.

Список литературы

[1] A. Apostolico, M. Farach, and C. S. Iliopoulos. Optimal superprimitivity testing for strings. *Inf. Process. Lett.*, 39(1):17–20, 1991.

[2] F. Blanchet-Sadri. Periodicity on partial words. *Computers and Mathematics with Applications*, 47(1):71–82, 2004.

[3] F. Blanchet-Sadri. Primitive partial words. *Discrete Applied Mathematics*, 148:195–213, 2005.

[4] F. Blanchet-Sadri and R.A. Hegstrom. Partial words and a theorem of Fine and Wilf revisited. *Theoretical Computer Science*, 270(1/2):401–419, 2002.

[5] S. Constantinescu and L. Ilie. Generalised Fine and Wilf's theorem for arbitrary number of periods. *Theor. Comput. Sci.*, 339(1):49–60, 2005.

[6] C. Corduneanu and H. Bohr. *Almost Periodic Functions.* Chelsea Publishing Company; 2nd edition, 1989.

[7] N. Fine and H. Wilf. Uniqueness theorems for periodic functions. *Proc. Amer. Math. Soc.*, 16:109–114, 1965.

[8] T. Harju. Defect theorem, lecture notes of "Combinatorics of words"Tarragona course, 2002/2003.

[9] T. Harju and J. Karhumäki. Many aspects of defect theorems. *Theor. Comput. Sci.*, 324(1):35–54, 2004.

[10] L. Boasson J. Berstel. Partial words and a theorem of Fine and Wilf. *Theor. Comput. Sci.*, 218(1):135–141, 1999.

[11] J. Karhumäki, F. Mignosi, and W. Plandowski. The expressibility of languages and relations by word equations. In *ICALP'97*, number 1256 in LNCS, pages 98–109. Springer-Verlag, 1997.

[12] G.O.H. Katona and D.O.H. Szász. Matching problems. *J. of Combinatorial Theory Ser B*, 10(1):60–92, 1971.

[13] Donald Knuth. *The Art of Computer Programming, Volume 4, Fascicle 3: Generating All Combinations and Partitions.* Addison-Wesley, 2005.

[14] A.M. Shur and Y.V. Konovalova. On the periods of partial words. In *MFCS'01*, number 2136 in LNCS, pages 657–665. Springer-Verlag, 2001.

[15] R.J. Simpson and R. Tijdeman. Multi-dimensional versions of a theorem of Fine and Wilf and a formula of Sylvester. *Proc. Amer. Math. Soc.*, 131:1661–1667, 2003.

[16] N.J.A. Sloane. Sequence A067824. The On-Line Encyclopedia of Integer Sequences, http://www.research.att.com/~njas/sequences/A067824.