УДК 519.16, 519.17

Ю.Г. СМЕТАНИН, М.В. УЛЬЯНОВ

РЕКОНСТРУКЦИЯ СЛОВ ПО КОНЕЧНОМУ МУЛЬТИМНОЖЕСТВУ ПОДСЛОВ В ГИПОТЕЗЕ СДВИГА 1. I. РЕКОНСТРУКЦИЯ БЕЗ ЗАПРЕТОВ[1]

Аннотация. Рассмотрена задача реконструкции слов по заданному множеству подслов в гипотезе, что оно порождено смещением окна фиксированной длины по неизвестному слову со сдвигом 1. Предложено решение для задачи реконструкции слов без запрещенного подслова, основанное на поиске эйлеровых путей или циклов в мультиорграфе де Брейна путем символического умножения матриц смежности с применением специальных операций умножения и сложения имен дуг. Рассмотрены особенности задачи и метод ее решения, позволяющий найти как число реконструкций, так и реконструируемые слова.

Ключевые слова: реконструкция слов, графы, эйлеровы пути, перечисление путей, реконструкция без запретов, графы де Брейна, символическое умножение матриц.

ВВЕДЕНИЕ. ОБЛАСТИ ПРИМЕНЕНИЯ

В настоящей работе рассматривается задача, которая по проблематике относится к комбинаторике слов — новому современному разделу дискретной математики.

Комбинаторика слов — термин, появившийся приблизительно тридцать лет назад для объединения направлений исследований, связанных общими подходами, но рассматриваемых в различных областях математики — от теории чисел до теоретической информатики. Главной целью исследований в этой области является изучение слов как самостоятельных объектов с точки зрения их внутренней структуры. Комбинаторика слов охватывает комбинаторные методы анализа множеств слов, используемые в теории формальных языков и автоматов [1], теории групп [2], теории хаоса [3], фрактальном анализе [4], символической динамике [5] и анализе временных рядов, биоинформатике [6] (в которой комбинаторные методы применял еще Г. Гамов [7]), лингвистике и др. Прогресс в данной области описывается в регулярно появляющихся книгах группы математиков, публикуемых под псевдонимом M. Lothaire [8–10].

Объектами рассмотрения в комбинаторике слов являются слова над произвольными алфавитами, а предметом исследований — изучение комбинаторных свойств различных множеств слов, как конечных, так и бесконечных. В реальных прикладных задачах информация о словах часто оказывается неполной; например, такая ситуация неизбежна в анализе бесконечных временных рядов, измеряемых на протяжении конечных интервалов времени.

В настоящей статье рассматривается одна из постановок задачи реконструкции слов по их известным последовательным фрагментам — подсловам. Задачи реконструкции слов тесно связаны, например, с задачами кодирования [11] и распознаванием образов [12], а также возникают в ряде других областей. Кодирование информации и распознавание образов — в некотором смысле предельные случаи задачи реконструкции слов: задачу построения кодов можно считать задачей определения множеств слов, восстанавливаемых по единственному искаженному образцу при искажениях заданного вида; задачи реконструкции с неформальным описанием классов слов относятся к классу задач распознавания.

В задачах исследования временных рядов [13] кодирование значений наблюдаемой величины может осуществляться в некотором алфавите, например (A, B, C, D, E, F), символами которого могут быть именованы полусегменты значений наблюдаемой величины в порядке их возрастания: A — имя полусегмента

[1] Работа выполнена при поддержке РФФИ, грант № 13-07-00516.

наименьших значений, F — наибольших. Если наблюдения ведутся в дискретном времени, то описание значений временного ряда по именам полусегментов есть слово над алфавитом имен. В случае, если наблюдаемый процесс характеризуется резкими выбросами значений наблюдаемой величины (до уровня F) относительно базального уровня (A, B) за один дискрет времени, так же, как и резкими спадами (от F до B), то получаемые кодовые слова временного ряда не будут содержать подслов CDE и EDC. Если при этом исходные данные — разрозненные фрагменты наблюдений, то задача реконструкции слова без запрещенных подслов является задачей восстановления всего описания временного ряда в предположении об особенностях его поведения.

В биоинформатике с комбинаторикой слов наиболее тесно связаны задачи анализа последовательностей ДНК [6, 14] и распознавания вторичной структуры белков [15]. Задача распознавания вторичной структуры белка заключается в следующем. Белок можно представлять как одномерную последовательность аминокислот или как одномерную последовательность характерных локальных конфигураций. В настоящее время общепринятым является допущение, что первичная структура однозначно определяет вторичную. При этом задача определения вторичной структуры (структуры локальных конфигураций) формулируется как задача преобразования слов в алфавите имен аминокислот в слова над алфавитом локальных конфигураций с помощью кодов скользящего блока.

При описании бизнес-процессов аппаратом теории графов [16] модель (граф бизнес-процесса) можно представить следующим образом: состояния процесса кодируются именованными вершинами, а переходы состояний — ребрами, отождествленными с этапами бизнес-процесса. Тогда запись конкретной реализации бизнес-процесса есть некоторое слово над алфавитом имен вершин, отражающее порядок перехода состояний. Если процесс физически распределен между различными организациями, то, скорее всего, получим информацию о его полном прохождении в виде набора подслов. При этом запрещенные подслова могут быть интерпретированы как нарушения модели — регламента бизнес-процесса. Возникающая задача реконструкции без запрещенных подслов содержательно означает возможность полной реконструкции всего процесса, соответствующего теоретической модели.

Таким образом, представляет интерес подробное изучение различных вариантов задачи реконструкции слов по некоторому множеству подслов меньшей длины, интерпретируемых как множество последовательных фрагментов неизвестного слова. При этом выделяется случай, когда реконструируемое слово не содержит заранее заданного запрещенного подслова. Один из возможных вариантов решения этой задачи на основе подслов фиксированной длины в гипотезе сдвига 1 и составляет предмет исследования настоящей работы; первая часть посвящена реконструкции без запрещенного слова; реконструкция при наличии запрета будет рассмотрена во второй части.

ТЕРМИНОЛОГИЯ И ОБОЗНАЧЕНИЯ

В статье используются терминология и обозначения как общепринятые в теории формальных грамматик и языков [1] и символической динамике [5], так и специальные авторские обозначения, связанные со спецификой изучаемой задачи. В теории языков исторически термины «слово» и «строка» считаются равноположенными. Например, известная задача символьного поиска формулируется как «Поиск подстроки в строке» [14]. В постановках задач реконструкции [12] более употребим термин «слово», который используется далее. Также, не теряя общности, будем рассматривать слова, порожденные бинарным алфавитом. Введем следующие обозначения:

- $\Sigma = \{0, 1\}$ — бинарный алфавит, s — произвольный символ алфавита;
- $\Sigma^0 = \varnothing$ — пустое множество;
- Σ^k — k-я декартова степень множества Σ (k-элементный кортеж);

• $\Sigma^* = \bigcup_{k=0}^{\infty} \Sigma^k$ — транзитивное замыкание Σ (множество всех возможных кортежей);

• c — произвольный кортеж (для кортежа c определим длину — число составляющих его элементов — как мощность мультимножества, образованного из кортежа: $|c| = n$, например $|(0,1,1,0)| = 4$; длина пустого кортежа равна нулю);

• $R(c, i)$ — функция выбора элемента кортежа, определенная при $1 \le i \le |c|$ и возвращающая элемент с номером i из кортежа c; если $i > |c|$, то $R(c, i) = \varnothing$; если элементы кортежа принадлежат Σ, то функция возвращает символ алфавита, например $R((1,0,0,1),3) = 0$;

• w — слово (над алфавитом) — последовательность символов алфавита, при этом собственно символы алфавита — слова по определению;

• λ — пустое (не содержащее символов) слово;

• $+$ — операция конкатенации (склейки) слов: $w_1 + w_2 = w_1 w_2$; результат операции $+$ — слово, представляющее собой последовательность символов слова w_1, за которой следует последовательность символов слова w_2, например $01+11 = 0110$;

• $D(\cdot)$: $D(c) = w$, где $c \in \Sigma^*$ — кортеж, w — слово; оператор $D(\cdot)$ — деструктор кортежа — оператор создания слова из кортежа путем конкатенации символов алфавита Σ,

$$D(c) = w = R(c,1) + R(c,2) + \ldots + R(c,|c|),$$

например $D((1,1,0,1)) = 1101$;

• $L(\cdot)$: $L(C) = W$, где $C \subseteq \Sigma^*$ — множество кортежей, W — множество слов; $L(\cdot)$ — оператор создания множества слов, состоящих из символов алфавита Σ, действующий на множество кортежей посредством последовательного применения оператора $D(\cdot)$,

$$L(C) = W = \{w \mid \forall c \in C w = D(c)\},$$

например $L(\{(1,1,0,1), (0,1,1)\}) = \{1101, 011\}$;

• $w = s_1 s_2 \ldots s_k \in L(\Sigma^k)$ — произвольное слово из множества $L(\Sigma^k)$ над алфавитом Σ;

• $|w| = k = |L^{-1}(w)|$ — длина слова, определяемая как число элементов в кортеже, порождающем это слово;

• $L_k = L(\Sigma^k) = \{w \mid |w| = k\}$ — множество всех слов длины k над алфавитом Σ;

• $L^* = L(\Sigma^*)$ — полный язык над алфавитом Σ — множество всех возможных слов;

• $L \subset L^*$ — произвольный неполный язык над алфавитом Σ.

Пусть $w = s_1 s_2 \ldots s_n \in L(\Sigma^n)$, тогда при $k < n$ имеем: $u = s_{i_1} s_{i_2} \ldots s_{i_k}$, $1 \le i_1 < i_2 < \ldots < i_k \le n$, — фрагмент слова w длины k, $v = s_{i_1} s_{i_2} \ldots s_{i_k}$, $1 \le i_1, i_2 = i_1 + 1, \ldots, i_k = i_{k-1} + 1 \le n$, — подслово слова w длины k; $Q(w, i, l)$ — оператор выделения подслова длины l в слове w, начиная с символа в позиции i. Пусть $|w| = n$, тогда оператор определен при $i + l - 1 \le n$:

$$Q(s_1 s_2 \ldots s_n, i, l) = u = s_i s_{i+1} \ldots s_{i+l-1}.$$

Для следующих трех операторов полагаем, что $|w| = k \ge 2$:

• $P(w) = Q(w, 1, k-1) = s_1 s_2 \ldots s_{k-1} \in L(\Sigma^{k-1})$ — полный префикс длины $|w| - 1$ слова w;

• $S(w) = Q(w, 2, k-1) = s_2 \ldots s_k \in L(\Sigma^{k-1})$ — полный суффикс длины $|w| - 1$ слова w;

• $Sn(w) = Q(w, k, 1) = s_k \in L(\Sigma^1)$ — суффикс слова w длины 1 — символ алфавита;

• $V(L_k, m)$ — оператор выборки: его результат — произвольное подмножество (возможно, с повторениями) из m слов множества L_k:

$$V(L_k, m) = \{v_i \mid i = \overline{1, m}; v_i = s_1^{(i)} s_2^{(i)} \dots s_k^{(i)} \in L_k\};$$

заметим, что в силу особенностей задачи допускаем рассмотрение $V(L_k, m)$ как мультимножества с кратностями элементов, в этом случае m — сумма кратностей элементов;

• $SH1(w, k)$ — оператор сдвига 1; определенный при $|w| > k$ оператор порождает множество подслов длины k мощности $|w| - k + 1$, выполняя сдвиг на 1 окна длины k по слову w, начиная с крайней левой позиции слова w:

$$SH1(w, k) = \{u_j \mid j = 1, |w| - k + 1; u_j = Q(w, j, k)\};$$

для оператора $SH1(w, k)$ допускаем создание мультимножества, например:

$$SH1(1101010, 4) = \{1101, 1010, 0101, 1010\} = \{1101, 1010^{(2)}, 0101\}.$$

ПОСТАНОВКА ЗАДАЧИ РЕКОНСТРУКЦИИ

Считаем заданными: длину подслова k, число подслов m и исходное мультимножество слов $V(L_k, m)$ над алфавитом $\Sigma = \{0, 1\}$, рассматриваемое как базис реконструкции. Принимаемая гипотеза сдвига 1 состоит в том, что $V(L_k, m)$ рассматривается как мультимножество подслов сдвига 1 относительно некоторого неизвестного слова w. В рамках данной задачи реконструкции исследуется постановка, в которой на реконструируемое слово не наложено дополнительных ограничений (запретов).

Постановка задачи реконструкции без запретов. Содержательно. Возможно ли в условиях гипотезы сдвига 1 относительно мультимножества $V(L_k, m)$ выполнить реконструкцию слова w в принципе, и если эта задача имеет решение, то является ли такая реконструкция единственной?

Формально. Введем в рассмотрение множество

$$W = \{w \mid |w| = m + k - 1,\ V(L_k, m) = SH1(w, k)\},$$

при этом равенство понимается как равенство мультимножеств (равны как элементы, так и их кратности). Тогда, если:

• $|W| = 0$ — решения нет, реконструкция невозможна и множество $V(L_k, m)$ не является реконструирующим мультимножеством;

• $|W| = 1$ — решение есть и единственно (реконструкция возможна и однозначна);

• $|W| \geq 2$ — существует несколько решений (реконструкция возможна и многозначна).

В последнем случае представляет интерес нахождение точного числа решений, т.е. значения $M = |W|$, так же, как и самих решений задачи — слов, составляющих множество W.

ИСТОРИЯ ВОПРОСА И РЕЗУЛЬТАТЫ

Среди задач реконструкции слов по частичной информации наибольшее внимание исследователей привлекла задача, когда имеющаяся информация представляет собой или набор фрагментов неизвестного слова, или набор подслов известной длины k, построенных в соответствии с определенными правилами; при отсутствии информации о парах «правило, фрагмент». Ниже приводятся известные в настоящее время результаты для разных вариантов постановки данной задачи.

Случай 1. Для неизвестного слова известной длины n известно мультимножество всех его фрагментов длины k.

Впервые эта задача была сформулирована в [17], затем формализована в [18], где была найдена верхняя оценка: при $k > n/2$ восстановление неизвестного слова возможно, и эта реконструкция однозначна. Для нижней границы результат [19], полученный одним из авторов статьи, состоит в том, что при $k < \log_2 n$ имеются слова, неразличимые по мультимножеству всех фрагментов длины k. Затем задача о реконструкции по мультимножеству фрагментов, полученных с помощью произвольного заданного набора правил, была сведена к решению системы диофантовых уравнений определенного вида [20, 21]. На основании этого результата уточнена нижняя оценка: для однозначности реконструкции необходимо (но недостаточно) условие

$$k > c \log_2 n,\ c = \log_2 (1+\sqrt{5}).$$

Следующий шаг был сделан в работе [22], где верхняя граница понизилась до $\Theta(\sqrt{n})$.

Случай 2. Для неизвестного слова известной длины n известно множество всех его фрагментов длины k.

В таком случае известные верхняя и нижняя границы совпадают [23]: для однозначной реконструкции произвольного слова необходимо и достаточно выполнения условия $k > \lfloor n/2 \rfloor + 1$. Реконструкция осуществляется за $k \mid N \mid$ операций, где N — мощность множества фрагментов.

Для задачи восстановления произвольного слова по мультимножеству подслов удалось найти следующий общий результат. Если построить орграф де Брейна [24], вершины которого помечены подсловами из мультимножества, а ребра соответствуют вершинам, полученным одна из другой сдвигом на 1, то решения задачи реконструкции соответствуют гамильтоновым путям в графе. Очевидный недостаток этого подхода связан с тем, что он сводит исходную задачу о реконструкции к *NP*-трудной задаче о гамильтоновом пути в графе. Таким образом, задача о реконструкции по подсловам в общей постановке за полиномиальное время остается открытой.

РЕКОНСТРУКЦИЯ СЛОВ БЕЗ ЗАПРЕТА В ГИПОТЕЗЕ СДВИГА 1

В основе предлагаемого решения задачи реконструкции в гипотезе сдвига 1 лежит построение специального мультиорграфа де Брейна $G = (D, H)$ [24], где D — множество вершин, а H — множество дуг. Предлагаемая разметка вершин и дуг G позволяет свести задачу реконструкции без запретов к задаче поиска всех эйлеровых путей (циклов), существенно менее трудоемкой, чем поиск гамильтонова пути в графе.

Мультиорграф $G = (D, H)$ строится по мультимножеству $V(L_k, m)$ следующим образом.

Построение вершин. Обозначим v_i, $i = \overline{1, m}$, элементы $V(L_k, m)$ — слова длины k, интерпретируемые как подслова сдвига 1 по неизвестному слову w. Образуем из полных префиксов $P(v_i)$ и полных суффиксов $S(v_i)$ всех слов v_i объединенное множество без повторений:

$$T = \bigcup_{i=1}^{m} P(v_i) \cup \bigcup_{i=1}^{m} S(v_i) = \{t_i \mid i = 1, \mid T \mid\};$$

очевидно, что $1 \le \mid T \mid \le 2m$.

Будем считать, что множество слов T порождает множество имен вершин. Введем в рассмотрение множество номеров вершин $I = \{i \mid i = 1, \mid T \mid\}$ и поставим во взаимно однозначное соответствие каждому элементу множества I элемент (подслово)

из множества T, образовав тем самым множество упорядоченных пар. Полученное множество $D \subset I \times T$ является множеством вершин мультиорграфа де Брейна:

$$D = \{d_i = (i, t_i) \mid i = \overline{1, |T|}\}.$$

Построение дуг. Для построения множества дуг введем в рассмотрение обычное (без повторений) множество Vp, построенное по мультимножеству $V(L_k, m)$. Пусть $|Vp| = n$, очевидно, что $n \le m$, тогда

$$Vp = \{vp_i, i = \overline{1, n}\}.$$

Введя обозначение $vp_i^{(r_i)}$ для слова vp_i кратности r_i в $V(L_k, m)$, получим представление

$$V(L_k, m) = \{vp_i^{(r_i)}, i = \overline{1, n}\}, \sum_{i=1}^{n} r_i = m. \quad (1)$$

Элементы множества дуг представим в виде упорядоченных пятерок, состоящих из начальной вершины, конечной вершины, символического имени дуги, кратности и значения:

$$h_i = (d_j, d_l, e_i, r_i, vp_i).$$

Тогда обобщенные дуги h_i мультиорграфа де Брейна строятся с помощью следующей процедуры.

Для всех слов $vp_i^{(r_i)}$, $i = \overline{1, n}$, из $V(L_k, m)$, записанных в представлении (1), выполнить такие операции:

1) определить префикс $P(\cdot)$ и суффикс $S(\cdot)$ для слова vp_i;

2) найти вершины графа де Брейна d_j, d_l: $\exists\, j, l : R(d_j, 2) = P(\cdot)$, $R(d_l, 2) = S(\cdot)$, имена которых совпадают с префиксом и суффиксом слова vp_i; существование таких вершин гарантировано по построению;

3) поставить в соответствие слову $vp_i^{(r_i)}$ дугу h_i с начальной вершиной d_j, конечной вершиной d_l, символическим именем e_i, кратностью r_i и значением слова vp_i:

$$vp_i^{(r_i)} \to h_i = (d_j, d_l, e_i, r_i, vp_i). \quad (2)$$

Заметим, что если $P(\cdot) = S(\cdot) = d_j$, то дуга (d_j, d_j) — петля; отметим также, что, поскольку $V(L_k, m)$ — мультимножество, формально $G = (D, E)$ — мультиорграф. В целях дальнейшей обработки графа будем считать, что от вершины d_j к вершине d_l идет дуга h_i, помеченная именем $e_i = R(h_i, 3)$, с кратностью r_i. Таким образом, мультиорграф G содержит n обобщенных дуг, имеющих совокупно общую кратность m.

На основе построенного мультиорграфа $G = (D, E)$ решение задачи существования реконструкции определяется следующей леммой.

Лемма 1 (о существовании). Если мультиорграф G не эйлеров, то $|W| = 0$, задача не имеет решения и множество $V(L_k, m)$ не является реконструирующим мультимножеством. Все возможные решения задачи реконструкции, дающие $|W| \ge 1$, соответствуют эйлеровым путям (эйлеровым циклам) в G с учетом кратности дуг.

Доказательство. По сути, доказательство очевидно: эйлеров путь или эйлеров цикл в G включает все дуги с символическими именами e_i, $i = \overline{1, n}$, с учетом их кратностей. В силу (1) и процедуры построения обобщенных дуг (2) этот путь (цикл) проходит по всем словам v_i из мультимножества $V(L_k, m)$, причем по каждому слову по одному разу в силу определения эйлерова пути или цикла. Согласно принципу построения множества вершин этот путь (цикл) проходит по подсловам некоторого слова со смещением 1. Таким образом, имеем

$$\exists\, w : V(L_k, m) = SH1(w, k).$$

Если эйлерова пути (цикла) в G нет, что можно элементарно проверить [25], то $|W|=0$. Если в G существует эйлеров путь и он единственный, то $|W|=1$. В случае эйлерова цикла реконструкцию можно начать, выбрав любую вершину цикла в качестве начальной. Поэтому если есть несколько путей или хотя бы один цикл, то возможно $|W|>1$, поскольку не обязательно все различные пути или пути, построенные, начиная с различных вершин цикла, приводят к различным реконструируемым словам. Лемма доказана.

В силу леммы 1 конструктивные решения для задачи реконструкции без запретов получают эйлеровыми путями или циклами в G. Заметим, что дуга h_i и ее символическое имя e_i имеют одинаковый номер i. В силу определения h_i (формула (2)) введем дополнительно функцию, возвращающую слово vp_i из кортежа h_i по символическому имени дуги e_i:

$$u(e_i)=vp_i.$$

Тогда решение задачи реконструкции получаем с использованием следующей леммы.

Лемма 2 (о реконструкции). Пусть эйлеров путь или цикл в мультиорграфе де Брейна задан кортежем обхода имен дуг $Ew=(e_{\pi(\cdot)}, e_{\pi(\cdot)}, \ldots, e_{\pi(\cdot)})$, где $\pi(\cdot)$ — перестановка индексов имен дуг с учетом кратностей, $|Ew|=m$. Тогда реконструируемое в гипотезе сдвига 1 слово w представляет собой конкатенацию:

$$w=u(R(Ew,1))+Sn(u(R(Ew,2)))+\ldots+Sn(u(R(Ew,m))).$$

Доказательство. Реконструкция корректна по построению мультиорграфа де Брейна и принятой гипотезе сдвига 1. Слово w начинается с подслова, содержащегося в первой дуге эйлерова пути (цикла). В силу определения функций R и u эта подстрока есть $u(R(Ew,1))$. К подслову добавляются суффиксы длины 1 от подслов, представленных дугами эйлерова пути (цикла) в порядке их обхода, но эти суффиксы определяются через $\underline{Sn}(u(R(Ew,i)))$ в силу определения функции Sn.

В случае, если ребра e_i, $i=\overline{1,n}$, с учетом кратностей образуют эйлеров цикл, у реконструируемого слова w совпадают префикс и суффикс длины $k-1$. Тогда решениями реконструкции являются все пути, полученные циклическими перестановками дуг эйлерова цикла. Лемма доказана.

Пример. Пусть исходные данные для реконструкции таковы: длина слова $k=2$, число слов $m=4$, $V(L_2,4)=\{00,01,10,11\}$.

Построим мультиорграф G. Вершины G: префиксы и суффиксы исходных подслов образуют множество $T=\{0,1\}$; таким образом, мультиорграф G имеет две вершины, $D=\{d_1=(1,0), d_2=(2,1)\}$. Дуги G: во множестве $V(L_2,4)$ нет повторяющихся элементов, поэтому множество Vp совпадает с $V(L_k,m)$, $|Vp|=4$, и G — орграф. В соответствии с (2) получаем множество ребер:

$$h_1=(d_1,d_1,e_1,1,00),\ h_2=(d_1,d_2,e_2,1,01),$$

$$h_3=(d_2,d_1,e_3,1,10),\ h_4=(d_2,d_2,e_4,1,11).$$

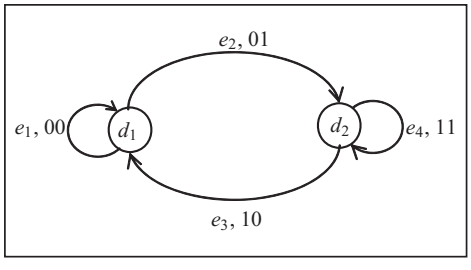

Рис. 1. Орграф де Брейна для множества подслов $V(L_2,4)=\{00,01,10,11\}$

Таким образом, G — орграф де Брейна с двумя вершинами, двумя дугами и двумя петлями. Орграф приведен на рис. 1, где для дуг показаны только символические имена и значения vp_i.

В соответствии с леммой о реконструкции для решения задачи реконструкции без запретов необходимо указать способ поиска и перечисления всех эйлеровых путей (циклов) в мультиорграфе де Брейна.

ПОИСК И ПЕРЕЧИСЛЕНИЕ ВСЕХ ЭЙЛЕРОВЫХ ПУТЕЙ

Для задачи поиска всех эйлеровых путей используем идею возведения в степень матрицы смежности графа на основе символического умножения имен дуг [25]. Считаем далее, что эйлеров цикл, при фиксации начальной вершины обхода цикла, является эйлеровым путем.

В соответствии с [25] рассмотрим матрицу смежности графа G — квадратную матрицу A размера $|D|$ и процедуру ее построения.

Инициализация: $a_{ij} = \varnothing \ \forall i, j = \overline{1, |D|}$.

Для всех дуг h_i, $i = \overline{1, n}$, из множества H выполнить следующие операции:

1) определить начальную и конечную вершины дуги — $d_j = R(h_i, 1)$, $d_l = R(h_i, 2)$;

2) присвоить элементу a_{jl} матрицы A значение в виде кортежа длины 1, элементом которого является символическое имя дуги h_i — $(e_i) = (R(h_i, 3))$.

На основе теоремы [25, с. 108] для определения маршрутов, состоящих из m дуг, необходимо возвести матрицу A в степень m и проанализировать полученные маршруты.

Особенности поиска эйлеровых путей в мультиорграфе заключаются в том, что в маршруте должны присутствовать все дуги в точном количестве их кратностей, тогда кортеж $Ew^{(m)} = (e_{\pi(\cdot)}, e_{\pi(\cdot)}, \ldots, e_{\pi(\cdot)})$ — эйлеров путь. В связи с этим вводим специальную операцию умножения имен дуг, используя информацию об их кратностях.

Определим содержательно операцию символического умножения $(*)$ кортежа имен на одноэлементный кортеж для получения элементов произведения $A^k * A \ \forall k = \overline{1, m-1}$ как операцию добавления имени дуги в кортеж при условии, что не превышена кратность данной дуги в исходном графе. Для кортежа символических имен $Ew^{(k)} = (e_{\pi(\cdot)}, e_{\pi(\cdot)}, \ldots, e_{\pi(\cdot)})$, $|Ew^{(k)}| = k$, определим дополнительно функцию $N(Ew^{(k)}, e_i)$, значением которой является кратность дуги e_i в кортеже $Ew^{(k)}$. Тогда операция символического умножения $(*)$ определяется следующим образом:

$$\begin{cases} Ew^{(k)} * (e_i) = \begin{cases} Ew^{(k+1)} = Ew^{(k)} \times (e_i), & N(Ew^{(k)}, e_i) + 1 \le R(h_i, 4), \\ \varnothing, & N(Ew^{(k)}, e_i) + 1 > R(h_i, 4), \end{cases} \\ Ew^{(k)} * \varnothing = \varnothing, \\ \varnothing * (e_i) = \varnothing, \\ \varnothing * \varnothing = \varnothing. \end{cases} \tag{3}$$

В результате допустимого в смысле (3) умножения получаем кортеж

$$Ew^{(k+1)} = Ew^{(k)} \times (e_i) = (e_{\pi(\cdot)}, e_{\pi(\cdot)}, \ldots, e_{\pi(\cdot)}, e_i),$$

в котором кратности дуг не превышают кратности дуг исходного мультиорграфа.

Определим операцию символического сложения $\oplus$, используемую при символическом умножении $A^k * A$, как операцию, обозначающую наличие нескольких кортежей в элементе $a_{ij} \in A^{k+1}$. Содержательно наличие нескольких кортежей в a_{ij} означает наличие нескольких путей из вершины d_i до вершины d_j, состоящих из $k+1$ дуг, т.е. путей длины $k+1$.

Лемма 3. Матрица A^m, где m — число дуг с учетом их кратностей в $G = (D, H)$, в непустых элементах содержит все эйлеровы пути графа G, при этом символическое умножение $(*)$ кортежа имен на одноэлементный кортеж выполняется по правилу (3), а операция символического сложения $\oplus$ означает наличие нескольких кортежей в элементе a_{ij}.

Доказательство. Эйлеров путь проходит по всем ребрам орграфа. Число прохождения через ребро равно кратности этого ребра. Согласно теореме [25, с. 108] матрица A^m содержит все маршруты, состоящие из m дуг, а в силу введенной операции умножения $(*)$ кратность каждой дуги в кортеже $Ew^{(m)}$ в точности равна их кратности в исходном мультиорграфе де Брейна. Лемма доказана.

Проиллюстрируем поиск эйлеровых путей на примере графа $G = (D, H)$, построенного по множеству подслов $V(L_2,4) = \{00, 01, 10, 11\}$ (рис. 1). Умножение выполнено по формуле (3). Отметим, например, что соответствующие элементы матрицы A^2 описывают все пути длины два между соответствующими вершинами без повторов дуг:

$$A = \begin{pmatrix} (e_1) & (e_2) \\ (e_3) & (e_4) \end{pmatrix},$$

$$A^2 = \begin{pmatrix} (e_2, e_3) & (e_1, e_2) \oplus (e_2, e_4) \\ (e_3, e_1) \oplus (e_4, e_3) & (e_3, e_2) \end{pmatrix},$$

$$A^3 = \begin{pmatrix} (e_2, e_3, e_1) \oplus (e_1, e_2, e_3) \oplus (e_2, e_4, e_3) & (e_1, e_2, e_4) \\ (e_4, e_3, e_1) & (e_3, e_1, e_2) \oplus (e_4, e_3, e_2) \oplus (e_3, e_2, e_4) \end{pmatrix},$$

$$A^4 = \begin{pmatrix} (e_2, e_4, e_3, e_1) \oplus (e_1, e_2, e_4, e_3) & \varnothing \\ \varnothing & (e_4, e_3, e_1, e_2) \oplus (e_3, e_1, e_2, e_4) \end{pmatrix}.$$

Таким образом, в данном графе существует эйлеров цикл. Предложенная процедура позволила получить все возможные пути, которые могут быть образованы из этого цикла. Напомним, что каждому пути соответствует некоторое восстанавливаемое по нему слово. Построим эти слова, используя правило реконструкции, указанное в лемме 2:

$$\begin{aligned} Ew &= (e_2, e_4, e_3, e_1) \Rightarrow w = 01100, \\ Ew &= (e_1, e_2, e_4, e_3) \Rightarrow w = 00110, \\ Ew &= (e_4, e_3, e_1, e_2) \Rightarrow w = 11001, \\ Ew &= (e_3, e_1, e_2, e_4) \Rightarrow w = 10011. \end{aligned} \tag{4}$$

В данном случае все слова различны, поэтому исходное множество $V(L_2,4) = \{00,01,10,11\}$, рассматриваемое как множество подслов в гипотезе сдвига 1, является реконструирующим, а порожденное им множество W состоит из четырех слов, указанных в (4).

ЗАКЛЮЧЕНИЕ

В работе предложен основанный на операторном подходе формализм описания операций над словами, порожденными конечным алфавитом. Исходя из этого, сформулирована постановка задачи реконструкции слов в гипотезе сдвига 1. Предложена процедура специальной разметки вершин и дуг в мультиорграфе де Брейна, позволившая свести решение задачи реконструкции к задаче поиска всех эйлеровых путей (циклов). В терминах описанного формализма сформулированы и доказаны леммы о существовании и реконструкции слов по подсловам в гипотезе сдвига 1. В результате адаптации символического умножения матриц к особенностям рассматриваемой задачи введена специальная операция символического умножения дуг, позволяющая строить эйлеровы пути в графе де Брейна.

Развитие данной постановки предполагает рассмотрение задачи реконструкции при наличии запрещенных слов, соответствующей реконструкции слов из заданного языка, а также исследование, анализ и обоснование вычислительной сложности алгоритма символического умножения, позволяющего определить все эйлеровы пути в графе де Брейна. Интерес к такому исследованию вызван некоторыми особенностями порождения дуг графа де Брейна, которые возникают в рассматриваемой задаче.

СПИСОК ЛИТЕРАТУРЫ

1. Хопкрофт Д., Мотвани Р., Ульман Дж. Введение в теорию автоматов, языков и вычислений. — М.: Изд. дом «Вильямс», 2008. — 528 с.
2. Morse M., Hedlund G. Unending chess, symbolic dynamics and a problem in semigroups // Duke Math. J. — 1944. — N 11. — P. 1–7.
3. Симиу Э. Хаотические переходы в детерминированных и стохастических системах. — М.: Физматлит, 2007. — 208 с.
4. Афраймович В., Угальде Э., Уриас Х. Фрактальные размерности для времен возвращения Пуанкаре. — Москва-Ижевск: Ин-т компьютер. исслед., R&C Dynamics, 2011. — 292 с.
5. Lind D., Marcus B. An introduction to symbolic dynamics and coding. — Cambridge (UK): Cambridge Univ. press, 1995. — 495 p.
6. Математические методы для анализа последовательностей ДНК. — М.: Мир, 1999. — 349 с.
7. Гамов Г. Комбинаторные принципы в генетике // Прикладная комбинаторная математика: Сб. статей / Под ред. Э. Беккенбаха. — М.: Мир, 1968. — С. 288–308.
8. Lothaire M. Combinatorics of words // Encyclopedia of mathematics and its applications. — Reading (Mass.): Addison-Wesley Publ. Co., 1983. — **17**. — 228 p.
9. Lothaire M. Algebraic combinatorics on words. — Cambridge: Cambridge Univ. press, 2002. — 455 p.
10. Lothaire M. Applied combinatorics on words // Encyclopedia of mathematics and its applications. — Cambridge: Cambridge Univ. press, 2005. — 610 p.
11. Левенштейн В.И. Восстановление объектов по минимальному числу искаженных образцов // Докл. РАН. — 1997. — **354**, № 5. — С. 593–596.
12. Леонтьев В.К. Распознавание двоичных слов по их фрагментам // Докл. РАН. — 1993. — **330**, № 4. — С. 434–436.
13. Dreyer W., Kotz Dittrich A., Schmidt D. Research perspectives for time series management systems // SIGMOD Record. — 1994. — **23**, N 1. — P. 10–15.
14. Гасфилд Д. Строки, деревья и последовательности в алгоритмах: Информатика и вычислительная биология / Пер с англ. — СПб.: Невск. диалект; БХВ-Петербург, 2003. — 654 с.
15. Рудаков К.В., Торшин И.Ю. Об отборе информативных значений признаков на базе критериев разрешимости в задаче распознавания вторичной структуры белка // ДАН. — 2011. — **441**, № 1. — С. 1–5.
16. Андерсен Б. Бизнес-процессы. Инструменты совершенствования / Пер. с англ. С.В. Ариничева; Науч. ред. Ю.П. Адлер. — М.: РИА «Стандарты и качество», 2003. — 272 с.
17. Калашник В.В. Восстановление слова по его фрагментам // Вычисл. математика и вычисл. техника. — Харьков, 1973. — Вып. 4. — С. 56–57.
18. Зенкин А.И., Леонтьев В.К. Об одной неклассической задаче распознавания // Журн. вычисл. математики и мат. физики. — 1984. — **24**, № 6. — С. 925–931.
19. Сметанин Ю.Г. Об алгебраической сложности задач восстановления векторов. — М., 1986. — Деп. в ВИНИТИ 03.09.86, № 6643–В86.
20. Леонтьев В.К., Сметанин Ю.Г. О восстановлении вектора по набору его фрагментов // Докл. АН СССР. — 1988. — **302**, № 6. — С. 1319–1322.
21. Leont'ev V.K., Smetanin Yu.G. Problems of information on the set of words // J. Math. Sci. — New York: Kluwer Acad./Consult. Bureau, 2000. — **108**, N 1. — P. 49–70.
22. Krasikov I., Rodity Y. On a reconstruction problem for sequences // J. Comput. Theory. Ser. A. — 1997. — **77**. — P. 344–348.
23. Леонтьев В.К. Задачи восстановления слов по фрагментам и их приложения // Дискрет. анализ и исслед. операций. — 1995. — **2**, № 2. — С. 26–48.
24. Bruijn N.G. de. A combinatorial problem // Indagationes Math. — 1946. — **8**. — P. 461–467.
25. Шапорев С.Д. Дискретная математика. Курс лекций. — СПб.: БХВ-Петербург, 2006. — 400 с.

Поступила 15.04.2013