ОБЩЕРОССИЙСКИЙ СЕМИНАР

«ИНФОРМАТИКА, УПРАВЛЕНИЕ И СИСТЕМНЫЙ АНАЛИЗ»

ВМК МГУ 27 октября 2015

О РЕКОНСТРУКЦИИ СЛОВ ПО ПОДСЛОВАМ В ГИПОТЕЗЕ СДВИГА 1

д. ф-м. н., г.н.с. ФИЦ ИУ РАН

СМЕТАНИН ЮРИЙ ГЕННАДИЕВИЧ

д.т.н., проф., в.н.с. ИПУ РАН

УЛЬЯНОВ МИХАИЛ ВАСИЛЬЕВИЧ
muljanov@mail.ru

МОСКВА

2015

ВВЕДЕНИЕ И ОБЛАСТИ ПРИМЕНЕНИЯ

По проблематике задача, рассматриваемая авторами, относится к комбинаторике слов — новому современному разделу дискретной математики. Объектами исследования в комбинаторике слов являются слова над произвольными алфавитами, а предметом исследований — изучение комбинаторных свойств различных множеств слов.

В настоящем докладе рассматривается одна из постановок задачи реконструкции слов по их известным последовательным фрагментам — подсловам. Задачи реконструкции слов оказываются тесно связанными:

— с задачами кодирования;

— с задачами распознавания образов;

— с задачами реконструкции длинных временных рядов;

— с необходимостью реконструкции бизнес-процессов по фрагментам их логов;

В биоинформатике задачи реконструкции слов наиболее тесно связаны с задачами анализа и восстановления последовательностей ДНК.

ТЕРМИНОЛОГИЯ И ОБОЗНАЧЕНИЯ (I)

Далее в докладе будет использоваться следующая терминология и обозначения:

Σ — алфавит, s — произвольный символ алфавита;

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

$Q(w, i, l)$ — 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;

ТЕРМИНОЛОГИЯ И ОБОЗНАЧЕНИЯ (II)

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

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

в силу особенностей задачи мы допускаем рассмотрение $V(L_k, m)$ как мультимножества с кратностями элементов;

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

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

Для оператора $SH1(w, k)$ мы, очевидно, допускаем создание мультимножества:

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

ПОСТАНОВКА ЗАДАЧИ РЕКОНСТРУКЦИИ (I)

Дано: длина подслова — k, число подслов — m, и исходное мультимножество подслов $V(L_k, m)$, рассматриваемое как базис реконструкции.

Гипотеза о мультимножестве $V(L_k, m)$ состоит в том, что мы рассматриваем его как мультимножество подслов сдвига 1 относительного некоторого неизвестного слова w.

I. Постановка задачи реконструкции без запретов

Содержательно: В условиях гипотезы сдвига 1 относительно мультимножества $V(L_k, m)$ возможно ли выполнить на основе подслов из $V(L_k, m)$ реконструкцию такого слова w, которое порождает это мультимножество окном сдвига 1?

Возможна ли такая реконструкция в принципе, и если эта задача имеет решение, то является ли такая реконструкция единственной?

ПОСТАНОВКА ЗАДАЧИ РЕКОНСТРУКЦИИ (II)

Формально: Введем в рассмотрение множество

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

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 многозначна).

В последнем случае очевидный интерес представляет получение точного числа решений, т.е. значения $M = |W|$, равно как и всех самих решений задачи — слов, составляющих множество W.

РЕКОНСТРУКЦИЯ ЭТАП 1 — МУЛЬТИОРГРАФ ДЕ БРЕЙНА (I)

В основе предлагаемого решения задачи реконструкции в гипотезе сдвига 1 лежит построение специального мультиорграфа де Брейна $G=(D,H)$, где D — множество вершин, а H — множество дуг. Предлагаемая авторами разметка вершин и дуг G позволяет свести постановку к задаче поиска всех эйлеровых путей или циклов — задаче, 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 через $v_i, i=\overline{1,m}$ 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 T (обычное множество без кратностей элементов):

$$T=\left(\bigcup_{i=1}^{m}P(v_i)\right)\cup\left(\bigcup_{i=1}^{m}S(v_i)\right)=\{t_i \mid i=1,|T|\},\qquad 1\le|T|\le 2m\,.$$

РЕКОНСТРУКЦИЯ ЭТАП 1 — МУЛЬТИОРГРАФ ДЕ БРЕЙНА (II)

Будем говорить, что множество слов T порождает множество имен вершин. Введем в рассмотрение множество номеров вершин $I = \{ i \mid i = 1, |T| \}$, и поставим во взаимнооднозначное соответствие каждому элементу множества I элемент (подслово) из множества T, образовав тем самым множество упорядоченных пар. Полученное множество $D \subset I \times T$ и есть множество вершин мультиорграфа де Брейна.

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

Построение дуг. Для построения множества дуг введем в рассмотрение обычное (без кратностей) множество Vp, построенное по мультимножеству $V(L_k, m)$. Пусть $|Vp| = n$, очевидно, что $n \le m$. Введем обозначение $vp_i^{(r_i)}$ для слова vp_i кратности r_i в $V(L_k, m)$, и получаем представление:

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

РЕКОНСТРУКЦИЯ ЭТАП 1 — МУЛЬТИОРГРАФ ДЕ БРЕЙНА (II)

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

$$h_i = \left(d_j, d_l, e_i, r_i, vp_i\right),$$

тогда обобщенные дуги h_i мультиорграфа де Брейна строятся следующей процедурой:

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

1.1. Определить префикс $P(\cdot)$ и суффикс $S(\cdot)$ для слова vp_i.

1.2. Найти вершины графа де Брейна $d_j, d_l : \exists j, l : R\left(d_j, 2\right) = P(\cdot), R\left(d_l, 2\right) = S(\cdot)$, имена которых совпадают с префиксом и суффиксом слова vp_i. Существование таких вершин гарантировано по построению. 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.

Заметим, что если $P(\cdot)=S(\cdot)=d_j$, то дуга (d_j,d_j) — петля; а также отметим тот факт, что поскольку $V(L_k,m)$ — мультимножество, то 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 с именем $e_i=R(h_i,3)$ и 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 доставляется 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 $|W|\geq 1$, соответствуют эйлеровым путям или эйлеровым циклам в G с учетом кратности дуг.

РЕКОНСТРУКЦИЯ ЭТАП 1 — МУЛЬТИОРГРАФ ДЕ БРЕЙНА (IV)

В силу леммы 1 конструктивные решения задачи даются эйлеровыми путями или циклами в G. Заметим, что дуга h_i и ее символическое имя e_i имеют одинаковый номер. Введем дополнительно функцию, возвращающую слово vp_i из кортежа h_i по символическому имени дуги e_i:

$$u(e_i) = vp_i,$$

тогда решения задачи реконструкции доставляется следующей леммой.

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:

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

ПРИМЕР ПОСТРОЕНИЯ ГРАФА де БРЕЙНА

Пусть длина слова $k = 2$, число слов $m = 4$, а исходное множество подслов

$$V(L_2,4) = \{00,01,10,11\}.$$

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. Получаем множество ребер:

$$h_1 = (d_1,d_1,e_1,1,00),\ h_2 = (d_1,d_2,e_2,1,01),\ h_3 = (d_2,d_1,e_3,1,10),\ h_4 = (d_2,d_2,e_4,1,11).$$

Таким образом, G— орграф де Брейна с двумя вершинами, двумя дугами и двумя петлями.

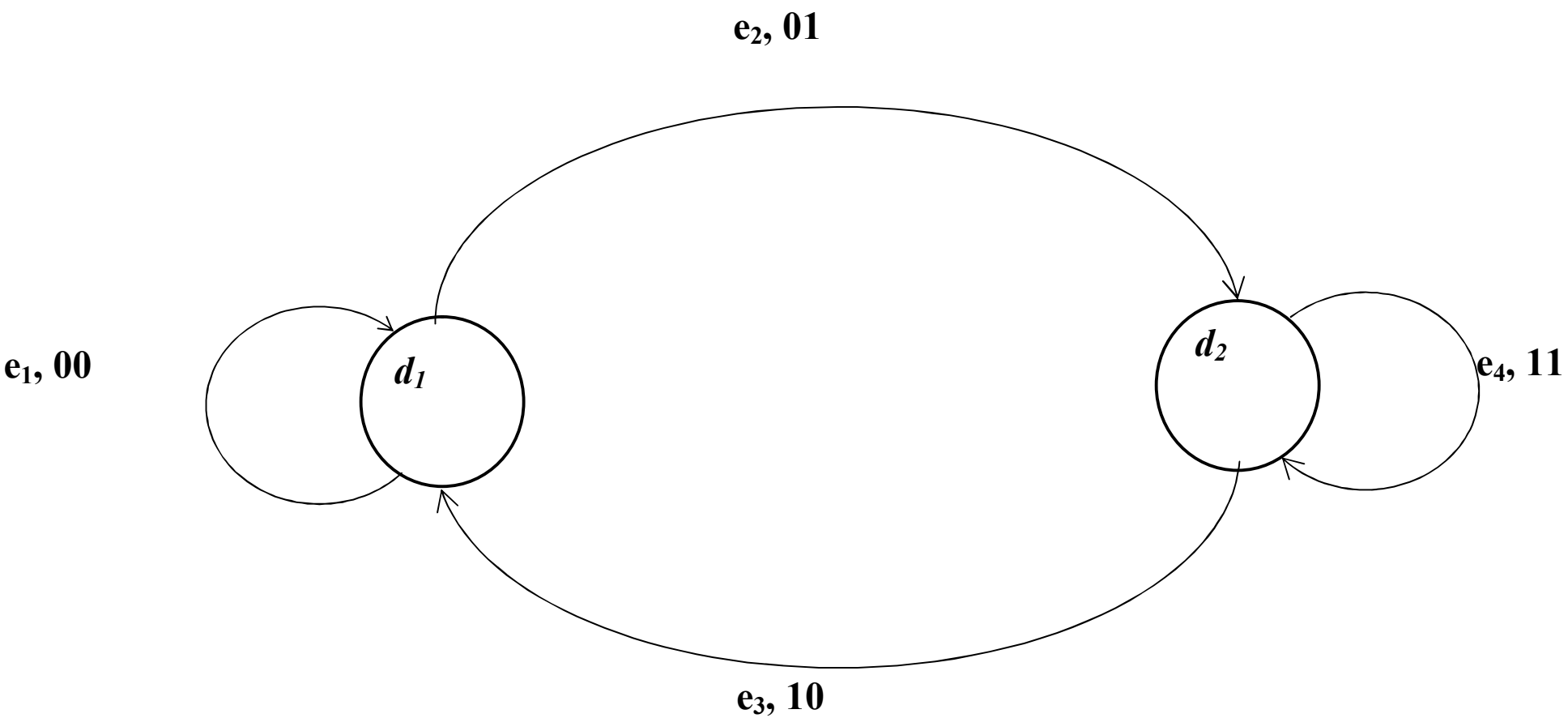

Рис. 1. Орграф де Брейна для множества подслов $V(L_2,4)=\{00,01,10,11\}$.

РЕКОНСТРУКЦИЯ ЭТАП 2 — ПЕРЕЧИСЛЕНИЕ ВСЕХ ЭЙЛЕРОВЫХ ПУТЕЙ (I)

Для задачи поиска всех эйлеровых путей воспользуемся методом возведения в степень матрицы смежности графа на основе специальной операции умножения символических имен дуг. Мы считаем далее, что эйлеров цикл, при фиксации начальной вершины, является эйлеровым путем.

Рассмотрим матрицу смежности графа G — квадратную матрицу A, размерности $|D|$ и следующую процедуру ее построения.

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

2. Для всех дуг $h_i, i = 1, n$ из множества H выполнить:

   2.2. Определить начальную и конечную вершину дуги — $d_j = R(h_i, 1)$, $d_l = R(h_i, 2)$.

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

РЕКОНСТРУКЦИЯ ЭТАП 2 — ПЕРЕЧИСЛЕНИЕ ВСЕХ ЭЙЛЕРОВЫХ ПУТЕЙ (II)

На основе известной теоремы о маршрутах, для определения числа маршрутов, состоящих из m дуг, необходимо возвести матрицу смежности в степень m.

Особенности поиска эйлеровых путей в нашем случае заключаются в том, что:

— элементы a_{jl} матрицы A содержат символические имена дуг;

— G есть мультиорграф, и в маршруте должны присутствовать все дуги в точном количестве их кратностей, тогда кортеж $Ew^{(m)} = \left(e_{\pi(\cdot)}, e_{\pi(\cdot)}, \ldots, e_{\pi(\cdot)}\right)$ — эйлеров путь.

В связи с этим мы вводим ***специальную алгебру умножения матриц, элементами которых являются символические имена дуг***, оперирующую дополнительно информацией о их кратностях.

1. Определим содержательно ***операцию символического умножения*** $(*)$ кортежа символических 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.

РЕКОНСТРУКЦИЯ ЭТАП 2 — ПЕРЕЧИСЛЕНИЕ ВСЕХ ЭЙЛЕРОВЫХ ПУТЕЙ (III)

Для кортежа символических имен $Ew^{(k)} = \left(e_{\pi(\cdot)}, e_{\pi(\cdot)}, \ldots, e_{\pi(\cdot)}\right), \left|Ew^{(k)}\right| = k$ определим дополнительно функцию $N\left(Ew^{(k)}, e_i\right)$, значением которой является кратность дуги e_i в кортеже $Ew^{(k)}$, тогда операция символического умножения $(*)$ определяется следующим образом:

$$\begin{cases} Ew^{(k)} * \left(e_i\right) = \begin{cases} Ew^{(k+1)} = Ew^{(k)} \times \left(e_i\right), \;\; N\left(Ew^{(k)}, e_i\right) + 1 \le R\left(h_i, 4\right); \\ \varnothing, \quad N\left(Ew^{(k)}, e_i\right) + 1 > R\left(h_i, 4\right); \end{cases} \\ Ew^{(k)} * \varnothing = \varnothing; \\ \varnothing * \left(e_i\right) = \varnothing; \\ \varnothing * \varnothing = \varnothing. \end{cases} \qquad (2)$$

В результате допустимого в смысле (2) умножения мы получаем кортеж:

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

в котором кратности дуг не превышают кратности дуг исходного мультиорграфа.

2. Определим ***операцию символического сложения*** «⊕», используемую при символическом умножении $A^k * A$ как операцию, размещающую объединяемые кортежи в одном элементе $a_{ij} \in A^{k+1}$. По сути это операция объединения кортежей в множество. 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 не пустых элементах $a_{ij} \in A^m$ содержит все эйлеровы пути графа G, при этом символическое умножение $(*)$ кортежа имен на одноэлементный кортеж выполняется по правилу (2), а операция символического сложения «⊕» порождает множество кортежей в элементах $a_{ij} \in A^k$, $k = 2, m$.

ПРИМЕР ПОИСКА ВСЕХ ЭЙЛЕРОВЫХ ПУТЕЙ

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

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

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

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

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

РЕКОНСТРУКЦИЯ ДЛЯ МОДЕЛЬНОГО ПРИМЕРА

Таким образом, в данном графе существует эйлеров цикл, и предложенное решение задачи позволяет получить все возможные пути, которые могут быть образованы из данного цикла с кратностью дуг 1. Напомним, что каждому пути соответствует некоторое восстанавливаемое по нему слово. Построим эти слова, используя правило реконструкции, указанное в лемме 2:

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

В данном случае все слова различны, и тем самым 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.

ОБОБЩЕНИЯ И ВАРИАНТЫ ЗАДАЧИ РЕКОНСТРУКЦИИ

Авторы рассмотривают дополнительно следующие возможные варианты и обобщения постановок задач реконструкции слов по подсловам:

1. Варианты задачи реконструкции

— реконструкция при наличии запретов;

(решение — статья авторов в журнале КиСА №1 2015 г.);

— реконструкция, содержащая априорно заданные подслова.

2. Обобщения задачи реконструкции

— реконструкция по подсловам переменной длины;

(предлагаемый вариант решения — сведение к реконструкции без запретов с последующей проверкой вхождения исходных «длинных» подслов, как априорно заданных)

— реконструкция по подсловам со сдвигом k;

— реконструкция двумерных слов.

Библиографический список

1. **Bruijn N.G. de.** A Combinatorial Problem // Nederl. Akad. Wetensch. Proc. 49. P. 758 – 764; Indagoniones Math. 1946. Vol. 8. P. 461–467.

2. **Lind D., Marcus B.** An Introduction to Symbolic Dynamics and Coding. Cambridge University Press, Cambridge, UK. 1995. — 495 pp.

3. **Lothaire M.** Algebraic Combinatorics on Words. 2005. — 610 c.

4. **Yu. G. Smetanin, M. V. Ulyanov** Reconstruction of a Word from a Finite Set of its Subwords under the unit Shift Hypothesis. I. Reconstruction without for Bidden Words // Cybernetics and Systems Analysis. January 2014, Volume 50, Issue 1, pp 148-156.

5. **Yu. G. Smetanin, M. V. Ulyanov** Reconstruction of a Word from a Finite Set of its Subwords Under the Unit Shift Hypothesis. II. Reconstruction with Forbidden Words // Cybernetics and Systems Analysis. January 2015, Volume 51, Issue 1, pp 157-164.

СПАСИБО ЗА ВНИМАНИЕ!