УДК 519.16, 519.17

Ю. Г. Сметанин, д-р. физ.-мат. наук, гл. науч. сотр., e-mail: ysmetanin@rambler.ru,
ФИЦ ИУ РАН (ВЦ РАН),
М. В. Ульянов, д-р. техн. наук, проф., e-mail: muljanov@mail.ru,
вед. науч. сотрудник, проф.,
ИПУ РАН им. В.А. Трапезникова, ВМК МГУ им. М.В. Ломоносова,
А.С. Пестова, магистрант,
НИУ ВШЭ

О числе возможных реконструкций слов по подсловам при окне переменного сдвига

Исходными объектами в данной статье являются конечные слова над бинарным алфавитом. Эти слова представляют собой символьные коды исследуемых объектов и процессов. В предположении о том, что исследователю известны лишь фрагменты (подслова) таких описаний, интерес представляет задача восстановления полного кода. С точки зрения комбинаторики слов восстановление описания на основе разрозненных фрагментов наблюдений представляет собой задачу реконструкции слова по известным подсловам. Для ее решения необходимо принять гипотезу о значении сдвига окна, порождающего данные подслова. Очевидно, что такая реконструкция может быть множественной. В статье предлагается оценка зависимости математического ожидания числа возможных реконструкций от значения параметра сдвига.

***Ключевые слова**: слова над конечным алфавитом, окно произвольного сдвига, реконструкция слов, оценка числа реконструкций*

Введение

Описание входной информации в терминах слов над фиксированным алфавитом возникает во многих задачах, в которых для исследования некоторых качественных свойств анализируемых систем или процессов можно использовать упрощенные символьные модели. Дискретизация во временной области и описание анализируемых систем словами, в которых символы соответствуют состояниям системы по квантам времени, дает возможность исследовать такие свойства, как периодичность, среднее время пребывания в определенных состояниях и т. д. без построения полной модели системы. Этот подход является предметом исследований в символической динамике [1]. Его реализация связана с решением ряда комбинаторных задач, чем вызвано появление комбинаторики слов как особой научной дисциплины, главная задача которой — изучение слов

как самостоятельного объекта с точки зрения их внутренней структуры [2]. Многие важные результаты, которые можно отнести к комбинаторике слов, были получены ранее в теории чисел, теории групп, теории вероятностей, кодировании. Исследования в комбинаторике слов часто связаны с взаимопроникновением методов из этих областей с целью создания новых подходов к решению задач.

Результаты реконструкции слов по частичной информации, полученные методами комбинаторики слов, эффективно используются при решении задач кодирования [3, 4, 5] и распознавания образов [6, 7, 8], задач биоинформатики, в том числе задач анализа последовательностей протеинов [9] и комбинаторного синтеза ДНК (геномной сборки) [10], а также при решении широкого круга задач символического анализа динамических систем [11].

Отметим, что кодирование и распознавание образов являются в некотором смысле предельными случаями реконструкции слов: задачу построения кодов можно считать задачей определения множеств слов, восстанавливаемых по единственному искаженному образцу при искажениях заданного вида, в то время как задачи реконструкции с неформальным описанием классов слов относятся к группе задач распознавания.

При решении задач символического анализа динамических систем возникающие символьные описания временных рядов, описывающие состояния системы по квантам времени, могут быть как неполными, так и искаженными и фрагментарными. При этом задача анализа временных рядов на основе разрозненных фрагментов наблюдений с точки зрения комбинаторики слов представляет собой задачу реконструкции слова по известными подсловам. Решение этой задачи позволяет восстановить описание временного ряда по наблюдаемым фрагментам.

Еще одной из возможных областей применения методов реконструкции слов является применение современных информационных технологий в анализе бизнес-процессов. При описании бизнес-процессов в терминах графов [12] состояния процесса кодируются именованными вершинами, а переходы состояний — ребрами, отождествленными с этапами бизнес-процесса. При этом запись конкретной реализации бизнес-процесса есть некоторое слово над

алфавитом имен вершин, отражающее порядок перехода состояний. Проблема реинжиниринга бизнес-процессов связана с задачами восстановления процесса по его известным фрагментам. Задача реконструкции бизнес-процесса возникает в связи с потерей или искажением информации о бизнес-процессе в целом, то есть потерей слов — описаний конкретных реализаций. Актуальность проблемы реинженириинга связана с быстрой адаптацией к часто меняющемуся рынку и выстраиванием наиболее эффективной работы организации. В этой проблематике аппарат комбинаторики слов также может быть эффективно использован для исследования бизнес-процессов [13].

В большинстве практических применений входные данные рассмотренных выше задач представляют собой фрагменты слов над некоторым конечным алфавитом. Основная проблема связана с фрагментарностью и неизвестным взаиморасположением фрагментов. Корректное описание такого взаиморасположения естественно приводит к рассмотрению окна со скользящим сдвигом. Движение такого окна вдоль неизвестного слова и порождает наблюдаемый набор подслов, соответствующих фрагментарному описанию. Именно в этих условиях и ставится задача реконструкции неизвестного слова, решение которой, как правило, приводит к множеству возможных реконструкций.

Решение задачи реконструкции при сдвиге один и фиксированной ширине окна предложено авторами в [14]. Обобщение этого решения на случай произвольного фиксированного сдвига основано на простой модификации предложенного метода и не представляет трудностей.

В прикладных задачах, как правило, значение сдвига окна неизвестно, в лучшем случае оно находится в некоторых пределах. Очевидно, что число возможных реконструкций при заданном наборе фрагментов зависит от предполагаемой величины сдвига: оно растет с увеличением сдвига, достигая факториального максимума при отсутствии перекрытий подслов (когда величина сдвига равна ширине окна).

Очевидный интерес представляет исследование зависимости числа возможных реконструкций от величины сдвига. В содержательном аспекте практических постановок задач чрезмерно большое число возможных

реконструкций, представляемых группе экспертов для принятия решения, приводит к резкому росту трудоемкости и времени принятия окончательного решения. В связи с этим экспертная группа заинтересована в получении ограниченного числа реконструкций, которые можно будет проанализировать за приемлемое время. Один из возможных путей сокращения предъявляемого числа реконструкций заключается во введении дополнительных ограничений, обусловленных характером прикладной задачи. В терминах комбинаторики слов такие ограничения приводят к задаче реконструкции с запретами, для которой авторами также предложено решение [15] в гипотезе единичного сдвига.сдвига один.

Примечание [41]: может быть, лучше сказать «в гипотезе единичного сдвига»? Да согласен!

В настоящей статье задача реконструкции слов при произвольном параметре сдвига рассматривается с целью оценки зависимости числа возможных реконструкций от величины сдвига при отсутствии запретов.

1. Терминология и обозначения

Введем следующие обозначения:

$\Sigma = \{a, b\}$ — бинарный алфавит, s — произвольный символ алфавита;
w — слово (над алфавитом Σ) — последовательность символов алфавита;

$|w|$ — длина слова (число символов в слове);

$SW(w,i,l)$ — оператор выделения подслова длины l в слове w, начиная с символа в позиции i. Пусть $|w| = r$, тогда оператор определен при $i + l - 1 \le r$:

Примечание [42]: здесь m-длина слова w, а ниже имеет другой смысл (см. прим.) Изменил в этом месте на r – это локальное обозначение в двух строчках

$$SW(s_1 s_2 \ldots s_r, i, l) = v = s_i s_{i+1} \ldots s_{i+l-1};$$

$SH(w,l,k)$ — оператор сдвига с параметром k, действующий на слово w окном ширины l и порождающий мультимножество подслов длины l с мощностью $\lfloor (|w| - l + k)/k \rfloor$. Определенный при $|w| \ge l$ оператор $SH(w,l,k)$ выполняет позиционирование окна ширины l последовательно, начиная с крайней левой позиции слова w, сдвигая каждый раз окно на k символов вправо по данному слову. В каждой позиции окна фиксируется подслово длины l:

$$SH(w,l,k) = \{ v_i \mid v_i = SW(w,i,l),\ i = 1, \lfloor (|w| - l + k)/k \rfloor, \Delta i = k \},$$

где подслова v_i могут, очевидно, порождать мультимножество.

Например, для алфавита $\Sigma = \{a,b\}$ и слова $w = "bbababa", |w| = 7$, при окне ширины 4 с параметром сдвига $k = 1$ имеем: $SH(bbababa,4,1) = \{bbab, baba, abab, baba\}$, а при параметре сдвига $k = 2$ получаем $SH(bbababa,4,2) = \{bbab, abab\}$, поскольку $\lfloor (|w| - 4 + 2)/2 \rfloor = 2$.

2. Постановка задачи

В целях формулировки задачи обозначим далее: исходное мультимножество подслов $V = \{v_i \mid v_i\ i = 1, n\}$, n — количество подслов во множестве V, l — длина подслова, k — параметр сдвига в операторе $SH(w,l,k)$, причем $1 \le k \le l$.

Объектом исследования в данной статье является мультимножество V подслов мощности n над бинарным алфавитом $\Sigma = \{a,b\}$. Слова в V имеют фиксированную длину l. Относительно мультимножества V мы принимаем гипотезу о том, что оно порождено оператором сдвига $SH(w,l,k)$ по неизвестным словам w. Последовательно принимаются гипотезы о фиксированном значении параметра сдвига k в операторе $SH(w,l,k)$ в диапазоне $1 \le k \le l$.

На основе мультимножества V возможно построение множества реконструируемых слов W при фиксированном параметре k. Полное решение задачи о реконструкции подробно описано в [14]. Очевидно, что возможное число реконструкций $|W|$ зависит как от параметров задачи (n, l и k), так и от собственно подслов в мультимножестве V. В предположении о том, что подслова в V порождены псевдослучайным равномерным генератором, введем в рассмотрение случайную величину $X(V,n,l,k) = |W|$ — число порождаемых реконструкций. Тем самым случайная величина $X(V,n,l,k)$ определена на вероятностном пространстве мультимножеств V мощности n, содержащих подслова длины l в гипотезе сдвига k. Обозначим через $N(n,l,k)$ оценку математического ожидания этой случайной величины — $N(n,l,k) = \hat{E}(X(n,l,k))$.

Предмет исследования — оценка математического ожидания числа реконструкций в зависимости от параметра k оператора сдвига $SH(w,l,k)$ при фиксированных значениях n и l на мультимножествах V, содержащих подслова бинарного алфавита в гипотезе равномерного распределения символов.

Содержательная постановка задачи

Постановка: для мультимножеств слов фиксированной длины в бинарном алфавите, оценить математическое ожидание числа возможных реконструкций с ростом параметра k оператора $SH(w,l,k)$ при фиксированных значениях n и l. По сути, мы принимаем гипотезу о том, что порожденные псевдослучайным равномерным генератором подслова в мультимножествах V представляют собой результат воздействия оператора $SH(w,l,k)$ на какие-то неизвестные слова w.

Формальная постановка задачи

Дано: мультимножество V подслов в бинарном алфавите, мощности n, все слова в котором имеют одинаковую длину l, полученных псевдослучайным равномерным генератором, и диапазон параметра сдвига k в операторе $SH(w,l,k)$ — $1 \le k \le l$.

Постановка: получить оценку математического ожидания числа реконструкций $N(n,l,k) = \hat{E}(X(n,l,k))$, в зависимости от параметра сдвига k, где случайная величина $X(V,n,l,k) = |W|$ есть число реконструкций, определяемое мультимножеством V при заданных параметрах n, l и k. Мы рассматриваем оценку $N(n,l,k)$ на множестве, элементами которого являются случайные мультимножества V при заданных параметрах n, l.

3. Оценка математического ожидания числа реконструкций

Введем в рассмотрение дополнительное обозначение $m = l - k$. Содержательно значение m есть длина перекрытия двух подслов при возможной реконструкции для порождающего окна со сдвигом k. Перед изложением предлагаемой оценки отметим следующее:

Примечание [43]: а здесь m - другой смысл, видимо, нужно здесь принять другое обозначение? Здесь m оставляем – оно везде дальше по тексту.

- авторы не рассматривают далее тривиальные и вырожденные случаи поставленной задачи (например, случай $k = 0$ и т.д.);
- поскольку дальнейший анализ выполняется аппаратом теории вероятностей, то получаемые значения $N(n,l,k)$ не обязательно являются целыми,

хотя в реальности одного испытания (реконструкции слов по множеству V) число возможных реконструкций есть или целое положительное число или ноль;

- возможное отсутствие реконструкций в терминах оценки математического ожидания означает близкие к нулю значения $N(n,l,k)$;
- при изменении параметра сдвига в операторе $SH(w,l,k)$ с фиксированными значениями n и l увеличивается длина L возможных реконструируемых слов, поскольку $L = l + (n-1)\cdot k$;
- как известно [16], математическое ожидание произведения двух независимых случайных величин X, Y равно произведению их математических ожиданий $E(X \cdot Y) = E(X) \cdot E(X)$, но в предлагаемом ниже решении требование независимости может нарушаться, что и приводит авторов к формулировке результата как оценки математического ожидания.

Если параметр k оператора $SH(w,l,k)$ равен длине слова, то мы получаем подслова без перекрытия, и, очевидно, возможными являются любые реконструкции, порожденные всеми перестановками подслов в V независимо от мощности алфавита. Тем самым при $k = l$ мы получаем точное решение задачи: $N(n,l,l) = n!$. Далее мы рассматриваем изменение параметра сдвига в пределах $1 \le k \le l-1$.

Будем считать процесс реконструкции последовательным, т.е. на очередном шаге к уже построенной части реконструируемого слова присоединяется если это возможно еще одно подслово из V, которое изымается из этого множества. Поскольку задача реконструкции ставится на всем мультимножестве V, то полное число шагов последовательной реконструкции равно мощности V, т.е. n. Идея построения оценки состоит в том, чтобы оценить число возможных присоединений на текущем шаге, и опираясь на свойство математического ожидания произведения получить требуемую оценку перемножением полученных оценок на всех шагах. Поскольку на каждом шаге присоединенное подслово изымается из V, мы не можем гарантировать независимость случайных величин, что, особенно заметно на последних шагах. Это приводит к формулировке результата как оценки математического ожидания.

Рассмотрим вначале случай, когда $k = l - 1$, т.е. окно оператора $SH(w,l,l-1)$ при сдвиге перекрывает только один последний символ предыдущего подслова. При этом $m = 1$. Поскольку процесс реконструкции последовательный, то на очередном шаге к уже построенной части слова присоединяется еще одно подслово из V, и процесс повторяется n раз до исчерпания множества V. Проведем следующее рассуждение, связанное с возможными вариантами последовательной реконструкции, оценивая последовательно математическое ожидание на каждом шаге:

● на шаге 1 в качестве начального подслова из V может быть взято любое подслово, и мы фиксируем n возможных вариантов;

● на шаге 2 в V осталось $n-1$ подслово. Поскольку мы рассматриваем бинарный алфавит, и порождение подслов в V происходит псевдослучайным равномерным генератором, то уже выбранное слово равновероятно оканчивается на один из двух символов алфавита, а в оставшейся части V подслова равновероятно имеют один из символов алфавита в качестве начального символа. Получаем, что оценка математического ожидания на этом шаге есть половина оставшихся подслов — $(n-1)/2$;

● на шаге j в V осталось уже $n-j+1$ подслов, и рассуждения, аналогичные приведенному выше показывают, что оценка математического ожидания на шаге j есть $(n-j)/2$;

● на последнем n-ом шаге в V осталось одно подслово, которое равновероятно либо присоединяется к уже построенной реконструкции, либо первый символ этого подслова не совпадает с последним символом текущей реконструкции, это дает значение оценки $1/2$.

Поскольку оценка математического ожидания числа возможных реконструкций равна произведению оценок математических ожиданий на шагах, мы получаем

$$N(n,l,l-1) = \frac{n!}{2^{(n-1)}}. \qquad (1)$$

Рассмотрим далее случай, когда $k = l - 2$, т.е. окно оператора $SH(w,l,l-2)$ при сдвиге перекрывает два последних символа предыдущего подслова, длина

перекрытия $m = 2$. Рассуждения аналогичные предыдущему, за исключением того, что на каждом шаге оценка математического ожидания числа претендентов на добавление к реконструкции равно четверти от текущей мощности множества подслов. Это очевидно, поскольку при фиксации двух символов в конце текущей реконструкции и в оставшейся части V нужное начало в подсловах встречается с вероятностью $1/4$. В результате мы получаем (при $m = 2$)

$$N(n,l,l-2) = \frac{n!}{4^{(n-1)}} = \frac{n!}{2^{2(n-1)}}. \quad (2)$$

Аналогичные рассуждения при произвольном допустимом $(1 \le k \le l-1)$ значении сдвига показывают, что в степень двойки в знаменателе формулы (2) входит длина перекрытия и число подслов — $m \cdot (n-1)$, но поскольку $m = l - k$, то мы, опираясь на (1) и (2) окончательно получаем

$$N(n,l,k) = \frac{n!}{2^{m(n-1)}} = \frac{n!}{2^{(l-k)(n-1)}}. \quad (3)$$

Простые преобразования, связанные с получением зависимости роста числа реконструкций от параметра сдвига при фиксированном числе подслов и их длине приводят к

$$N(n,l,k) = \frac{n!}{2^{l\cdot(n-1)}} 2^{k\cdot(n-1)}. \quad (4)$$

Таким образом, с увеличением параметра сдвига число реконструкций растет экспоненциально с основанием $2^{(n-1)}$. Проверим случай, когда параметр сдвига равен длине подслова $k = l$. Подставляя в (3) получаем

$$N(n,l,l) = \frac{n!}{2^{(l-l)(n-1)}} = \frac{n!}{2^0} = n!,$$

что согласуется с ранее полученным точным решением.

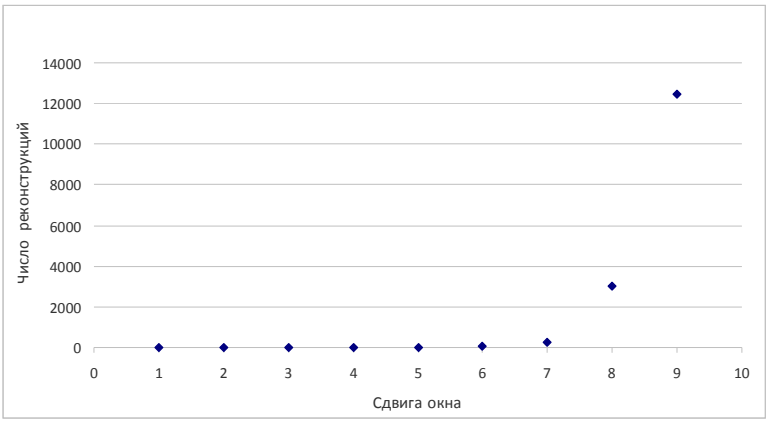
4. Экспериментальные результаты

Для проведения экспериментального исследования была разработана программа на языке Java, реализующая методику получения полного числа реконструкций, основанная на статье [14] (программа и экспериментальное исследование выполнено А.С. Пестовой). Отметим, что такое исследование достаточно трудоемко, поскольку результат символьного умножения матрицы

смежности мультиорграфа де Брейна хранит в виде символьного описания кортежей все эйлеровы пути в данном мультиорграфе [14]. Очевидно, что экспоненциальный рост числа реконструкций приводит к такой же сложности в программной реализации и соответствующим временным затратам. В связи с этим эксперименты проводились при количестве подслов $n=10$ и длине подслова $l=10$ с изменением сдвига от 1 до 10. Были получены следующие результаты (см. табл. 1 и рис.1).

Экспериментальное число реконструкций при $n=10, l=10$

k	1	2	3	4	5	6	7	8	9	10
$N(10,10,k)$	0	0	1	2	10	63	240	3040	12474	3628800



Экспериментальные результаты реконструкции — значения $N(10,10,k)$.

Результаты подтверждают экспоненциальный рост числа реконструкций, но расходятся численно с оценкой (4). Такое расхождение, по мнению авторов, связано с тем, что оценка математического ожидания числа реконструкций правомерна при больших значениях n, когда предположение о равной частотной встречаемости символов бинарного алфавита согласуется с исходными данными. При длине подслов в 10 символов и числе подслов в нашем исследовании (10) такое предположение вносит погрешность в оценку, поскольку для биномиального распределения при $n=10$ вероятность того, что в последовательности из десяти символов встречается пять символов «а» равна 0.246, четыре — 0.205, три — 0.117. Такие неравномерности распределения

частотной встречаемости символов приводят к сокращению числа возможных реконструкций.

Заключение

В статье получена оценка математического ожидания числа возможных реконструкций по исходному мультимножеству подслов в бинарном алфавите в зависимости от параметра сдвига порождающего окна при отсутствии запретов.

Полученные оценки дают возможность оценить связь между «плотностью» наличествующей фрагментарной информации и числом альтернативных решений. Они полезны с точки зрения возможности определения числа альтернатив до начала реконструкции, что может оказаться важным в задачах принятия решений на основе восстановленной полной информации.

Возможные дальнейшие исследования связаны с анализом влияния дополнительной информации в виде запретов некоторых вариантов реконструкции на число возможных реконструкций.

Список литературы

1. **Lind D., Marcus B.** An Introduction to Symbolic Dynamics and Coding. New York: Cambridge University Press, Cambridge, UK, 1995. 495 pp.
2. **Lothaire M.** Algebraic Combinatorics on Words. Cambridge University Press, 2002. 455 с. URL: http://www-igm.univ-mlv.fr/~berstel/Lothaire/.
3. **Мак-Вильямс Ф. Дж., Слоэн Н. Дж.** Теория кодов, исправляющих ошибки. М.: Связь, 1979. 744 с.
4. **Левенштейн В. И.** Двоичные коды с исправлением выпадений, вставок и замещений символов // Докл. АН СССР. 1965. Т. 163, № 4. С. 707 – 710.
5. **Левенштейн В. И.** Восстановление объектов по минимальному числу искаженных образцов // Докл. РАН. 1997. Т. 354, № 5. С. 593 – 596.
6. **Зенкин А. И., Леонтьев В. К.** Об одной неклассической задаче распознавания // Журн. вычисл. математики и матем. физики. 1984. Т. 24, № 6. С. 925 – 931.
7. **Леонтьев В. К.** Распознавание двоичных слов по их фрагментам // Докл. РАН. 1993. Т. 330. № 4. С. 434 – 436.
8. **Apostolico A., Atallah M. J.** Compact Recognizers of Episode Sequences // Information and Computation. 2002. Vol. 174. С. 180 – 192.
9. **Wang J. T. L., Ma Q., Shasha D., Wu C. H.** New Techniques for Extracting Features from Protein Sequences // IBM Systems Journal. 2001. Vol. 40, № 2. P. 426 – 441.

10. **Гасфилд Д.** Строки, деревья и последовательности в алгоритмах: Информатика и вычислительная биология / Пер. с англ. И. В. Романовского. СПб.: Невский Диалект, 2003. 654 с.

11. **Kortelainen J.** On the System of Word Equations in a Free Monoid // Journal of Automata, Languages and Combinatorics. 1998. Vol. 3, № 1. P. 43 – 57.

12. **Андерсен Бьёрн** Бизнес-процессы. Инструменты совершенствования / Пер. с англ. С.В. Ариничева. М.: РИА «Стандарты и качество», 2003. 272 с.

13. **Евдокимов А .А., Левин А. А.** Инструментарий графического исследования символьных последовательностей // Прикладная дискретная математика. 2008. Вып.1 (Прикладная теория графов).

14. **Smetanin Yu. G., Ulyanov M. V.** 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. Vol. 50, Iss. 1. P. 148-156.

15. **Smetanin Yu. G., Ulyanov M. V.** 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. Vol. 51, Iss. 1. P. 157-164.

16. **Ширяев. А. Н.** Вероятность. М.: МЦНМО, 2007. 968 с.

Yu. G. Smetanin, D. Sc., Chief Researcher, e-mail: ysmetanin@rambler.ru,

Federal Research Center "Informatics and Control", Moscow, Russia,
M. V. Uljanov, Ph.D., Leading Researcher, e-mail: muljanov@mail.ru,
Institute of Control Sciences, Moscow, Russia,
A. S. Pestova, Graduate Student,
National Research University Higher School of Economics, Москва, Russia

On The Number of Possible Reconstructions of Words Using Subwords with Windows of Different Shift

The objects studied in the article are finite words over binary alphabet. These words are symbolic codes of the objects and processes under study. Under the assumption that only fragments (subwords) of such descriptions are known to the researcher, the problem of reconstructing the complete code is of interest. From the point of view of combinatorics of words, reconstructing the description using disjointed fragments of observations is the task of reconstructing a word using known subwords. To solve it, one needs to accept the hypothesis about the value of the sliding window that generates the subwords. Obviously, such a reconstruction can be multivalued. The dependence of the mathematical expectation for the number of possible reconstructions on the value of the shift parameter is proposed in the article. The obtained estimates make can be used for evaluating the relationship between the "density" of the available fragmentary information and the number of alternative solutions of the reconstruction problem. They are useful from the point of the possibility of determining the number of alternatives before the beginning of the reconstruction process. The estimations may prove to be important in decision-making problems which demand the restored complete information. Possible further studies are related to the analysis of the effect of additional information, presented in the form of prohibitions of certain reconstruction options, on the number of possible reconstructions.

Keywords: *words over a finite alphabet, sliding window of arbitrary shift, reconstruction of words, estimation of the number of reconstructions*

References

1. **Lind D., Marcus B.** An Introduction to Symbolic Dynamics and Coding, New York, Cambridge University Press, Cambridge, UK. 1995, 495 p.
2. **Lothaire M.** Algebraic Combinatorics on Words, Cambridge University Press, 2002, 455 p. http://www-igm.univ-mlv.fr/~berstel/Lothaire/.
3. **Mcwilliams F.J. and Sloan N.J.A.** The Theory of Error Correcting Codes, North Holland, 1977.
4. **Levenshtein V.I.** *Dvoichnye kody s ispravlenied vypadenii, vstavok i zameshchenii simvolov* (Binary codes correcting insertions, deletions and substitutions, *Dokl. AN SSSR,* 1965, vol. 163, no. 4, pp. 707 – 710 (in Russian).

5. **Levenshtein V.I.** *Vosstanovlenie ob'ektov po minimal'nomu chislu iskazhennykh obraztsov* (Reconstruction of objects using the minimal number of corrupted samples), *Dokl. AN SSSR,*1974, vol. 354, no. 5, pp. 593 – 595 (in Russian).

6. **Zenkin A.I., Leont'ev V.K.** *Ob odnoi neklassicheskoi zadache raspoznavaniia* (On a non-classic recognition problem), *Zhurn. Vychisl. Matematiki i Matem. Fiziki,* 1984, vol. 24, no. 6, pp. 925 – 931 (in Russian).

7. **Leont'ev V.K.** *Raspoznavanie dvoichnykh slov po ikh fragmentam* (Binary words recognition using their fragments), *Dokl. RAN,* vol. 330, no. 4, pp. 434 – 436 (in Russian).

8. **Apostolico A., Atallah M. J.** Compact Recognizers of Episode Sequences, *Information and Computation, 2002, vol. 174, pp. 180 – 192*.

9. **Wang J. T. L., Ma Q., Shasha D., Wu C. H.** New Techniques for Extracting Features from Protein Sequences, IBM Systems Journal, *2001, vol. 40, no. 2, pp. 426 – 441*.

10. **Gasfield D.** Lines, Trees and Sequences in Algorithms, Informatics and Computational Biology, NEvsky Dialect, Petersburg, 2003, 654 p.

11. **Kortelainen J.** On the System of Word Equations in a Free Monoid, Journal of Automata, *Languages and Combinatorics, 1998, vol. 3, no. 1, pp. 43 – 57*.

12. **Andersen B.** Business Process Improvement Toolbox, ASQ Quality Press, 296 p. (2003)

13. **Evdokimov A.A., Levin A.A.** *Instrumentarii graficheskogo issledovaniia simvol'nykh posledovatel'nostei* (A tool for graphical exploration of character sequences), in: Evdokimov A.A., Levin A.A.: Novosibirsk: *Prikladnaia diskretnaia matematika* (Applied discrete vathematics), 2008, no. 1, *Prikladnaia teoriia grafov* (Applied graph theory).

14. **Smetanin Yu. G., Ulyanov M. V.,** Reconstruction of a Word from a Finite Set of its Subwords Under the Unit Shift Hypothesis. I. Reconstruction with Forbidden Words, *Cybernetics and Systems Analysis*, January 2014, vol. 51,issue 1, pp. 168-177.

15. **Smetanin Yu. G., Ulyanov M. V.,** 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.

16. **Shiryaev A. N.** *Veroyatnost'* (Probability), Moscow, MCNMO, 2007, 968 p.