DOI: 10.17323/2587-814X.2020.2.84.92

О возможности определения префикса и суффикса слова по подсловам фиксированной длины

Г.Н. Жукова[a]
E-mail: galinanzhukova@gmail.com

Ю.Г. Сметанин[b]
E-mail: smetanin.iury2011@yandex.ru

М.В. Ульянов[c]
E-mail: muljanov@mail.ru

[a] Национальный исследовательский университет «Высшая школа экономики»
Адрес: 101000, г. Москва, ул. Мясницкая, д. 20

[b] Федеральный исследовательский центр «Информатика и управление» Российской академии наук
Адрес: 119333, г. Москва, ул. Вавилова, д. 40

[c] Институт проблем управления им. В.А. Трапезникова Российской академии наук
Адрес: 117997, г. Москва, ул. Профсоюзная, д. 65

Аннотация

В прикладных задачах бизнес-информатики, связанных с анализом данных (в частности, при анализе и прогнозировании временных рядов, при исследовании лог-файлов бизнес-процессов) возникают задачи качественного анализа. Методы качественного анализа достаточно часто используют символьное кодирование как способ представления информации об исследуемых процессах. В ряде ситуаций, обусловленных фрагментарностью таких описаний, возникает задача реконструкции полного символьного описания процесса (слова) по его последовательным фрагментам (подсловам). По мультимножеству всех подслов достаточно большой длины исходное слово восстанавливается однозначно. В случае недостаточно длинных подслов возможно множество различных реконструкций исходного неизвестного слова. Число допустимых реконструкций можно сократить, если определить суффикс и префикс реконструируемого слова. Предложен метод определения префикса и суффикса слова над конечным алфавитом, состоящих из $k-1$ символов каждый, на основании мультимножества V подслов фиксированной длины, равной k. Принимается гипотеза о том, что это мультимножество порождено смещением на один символ окна фиксированной длины k по неизвестному слову. Метод определения префикса и суффикса основан на построении и анализе матрицы, образованной записанными по строкам в произвольном порядке подсловам из V и использовании оператора, действующего на мультимножества символов алфавита, образованных соседними столбцами этой матрицы. Метод позволяет определить префикс $a_1a_2 \ldots a_{k-1}$ и суффикс $b_1b_2 \ldots b_{k-1}$ неизвестного слова в случае, если $a_i \neq b_i$ для любых i от 1 до $k-1$. В случае, если $a_i \neq b_i$ только для некоторых значений i, в префиксе и суффиксе определяются символы в соответствующих позициях, а для остальных символов выполняется условие $a_j = b_j$. В худшем случае метод констатирует, что $a_i = b_i$ для всех i от 1 до $k-1$, но не определяет сами символы. Это ситуация, при которой префикс и суффикс совпадают, но не могут быть определены.

Ключевые слова: реконструкция слова; префикс; суффикс; мультимножество подслов; подслова фиксированной длины; оператор сдвига.

Цитирование: Жукова Г.Н., Сметанин Ю.Г., Ульянов М.В. О возможности определения префикса и суффикса слова по подсловам фиксированной длины // Бизнес-информатика. 2020. Т. 14. № 2. С. 84–92.
DOI: 10.17323/2587-814X.2020.2.84.92

Введение

В прикладных областях бизнес-информатики, связанных с анализом данных, таких как анализ и прогнозирование временных рядов [1–6], исследование лог-файлов бизнес-процессов [7] и др. возникают задачи качественного анализа. В этом случае одним из часто используемых способов представления информации о процессах является символьное кодирование [8]. При этом описание поведения временного ряда или бизнес-процесса кодируется словом над конечным алфавитом, которое и является объектом дальнейшего исследования. Однако в ряде случаев, в том числе при анализе бизнес-процессов и временных рядов, исследователи получают не само слово целиком, а множество подслов, которые являются последовательными фрагментами некоторого слова. Поскольку при этом позиции подслов в исходном слове неизвестны, возникает задача реконструкции – восстановления неизвестного слова по исходному множеству подслов [9–17]. Эта задача содержательно относится к специальному разделу дискретной математики – комбинаторике слов [18]. Объектами исследования 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 при анализе бесконечных временных рядов, измеряемых на протяжении конечных интервалов времени.

Заметим, что одной из важных областей практического применения методов комбинаторики слов является область биомолекулярных моделей и процессов. При этом работа с фрагментарной информацией характерна для ряда задач биоинформатики и геномики. Например, задача секвенирования геномов [19, 20] по сути является задачей реконструкции слов в условиях сильных ограничений, подразумевающей однозначность реконструкции.

Задачи восстановления слов над конечным алфавитом имеют различные постановки, отличающиеся как объемом имеющейся информацией, так и ограничениями на допустимые решения [21–23]. Обычно эти задачи, как задачи с неполной информацией, являются сложными, и получение какой-либо дополнительной информации, очевидно, позволяет сократить рассматриваемое множество возможных решений.

При качественном анализе временных рядов [24, 25] 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 наименьших значений, F – наибольших. Поскольку фиксация наблюдений ведется в дискретном времени, описание значений временного ряда по именам полусегментов есть слово над алфавитом имен. Если наблюдаемый процесс характеризуется резкими выбросами значений наблюдаемой величины (до уровня F) относительно базального уровня (A, B) за один дискрет времени, равно 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 представляют собой подслова – разрозненные фрагменты наблюдений, то задача реконструкции слова по подсловам есть задача восстановления всего описания временного ряда в предположении об особенностях его поведения.

Аналогичная ситуация возникает при реконструкции лог-файлов бизнес-процессов при наличии фрагментарной информации. При описании бизнес-процессов аппаратом теории графов [7] модель (граф бизнес-процесса) может быть представлена 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 и исполнителями, то, скорее всего, мы 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 интерес представляет как случай, когда реконструируемое слово не содержит заранее заданного запрещенного подслова, так и случай с наличием запрещенных подслов. Один из возможных вариантов решения этой задачи на основе подслов фиксированной длины в гипотезе сдвига один предложен в работах [26, 27]. Однако множество возможных реконструкций может быть достаточно велико и возникает задача о возможном сокращении числа претендентов на «правильное» реконструируемое слово. Мы хотим получить дополнительную информацию из исходного множества подслов, которая будет полезна при редукции полученного множества реконструкций. Речь идет о возможности восстановления и/или определения шаблона префикса и суффикса неизвестного слова, что в рамках процедуры редукции приведет к рассмотрению только тех слов, которые обладают полученными шаблонами префикса и суффикса. Именно эта задача и является предметом настоящей статьи.

1. Терминология и обозначения

Далее в тексте статьи будут использоваться следующие обозначения:

$\Sigma = \{s_1, s_2, \ldots s_l\}$ – алфавит, s_i – i-ый символ алфавита;

Σ^k – k-я декартова степень множества Σ (множество k-элементных кортежей);

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

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

$L(\cdot) : L(C) = W$, где $C \subseteq \Sigma^*$ – множество кортежей, W – множество слов. Оператор $L(\cdot)$ есть оператор создания множества слов, состоящих из символов алфавита Σ, действующий на множество кортежей;

a_i – i-ый символ слова w, $a_i \in \Sigma$;

$w = a_1 a_2 \ldots a_n \in L\left(\Sigma^n\right)$ – произвольное слово из n символов над алфавитом Σ;

$|w| = n$ – длина слова, определяемая как число элементов в порождающем кортеже;

$L_k = L\left(\Sigma^k\right) = \left\{w \,|\, |w| = k\right\}$ – множество всех слов длины k над алфавитом Σ.

Пусть $w = a_1 a_2 \ldots a_n \in L\left(\Sigma^n\right)$, тогда при $k < n$:

$v = a_{i_1} a_{i_2} \ldots a_{i_k}, 1 \le i_1, i_2 = i_1 + 1, i_k = i_{k-1} + 1 \le n$ – подслово слова w длины k;

$Q(w, i, k)$ – оператор выделения подслова длины k в слове w, начиная с символа в позиции i. Пусть $|w| = n$, тогда оператор определен при $i + k - 1 \le n$

$$Q(a_1 a_2 \ldots a_n, i, k) = a_i a_{i+1} \ldots a_{i+k-1},$$

$$Q(w, i, k) \in L_k;$$

Для следующих двух операторов полагаем, что $|w| = n \ge 2$ и $1 \le k < n$:

$P(w, k) = Q(w, 1, k) = a_1 a_2 \ldots a_k \in L_k$ – префикс длины k слова w;

$S(w, k) = Q(w, n - k + 1, k) = a_{n-k+1} \ldots a_n \in L_k$ – суффикс длины k слова w;

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

$SH1(w, k) = \{v_j | j = 1, |w| - k + 1; v_j = Q(w, i, k)\}$.

2. Постановка задачи

В дальнейшем мы считаем заданными: длину подслова – k, число подслов – m, а также исходное мультимножество подслов V над алфавитом Σ, рассматриваемое как базис реконструкции некоторого неизвестного слова w:

$$V = \left\{v_i \,|\, i = \overline{1,m}; v_i = a_{i1} a_{i2} \ldots a_{ik} \in L_k\right\}.$$

Принимаемая авторами гипотеза сдвига один состоит в том, что мы рассматриваем V как мультимножество подслов сдвига один относительного некоторого неизвестного слова w, при этом $|w| = n = m + k - 1$:

$$V = SH1(w,k) = \left\{ v_j \mid j = 1, n-k+1; v_j = Q(w,j,k) \right\}.$$

Содержательная постановка: В условиях гипотезы сдвига один относительно мультимножества V возможно ли определить префикс и суффикс длины $k-1$ неизвестного слова w, или получить какую-либо содержательную информацию о его префиксе и суффиксе?

Математическая постановка: По данному мультимножеству V с длиной подслова k и числом подслов m определить префикс $P(w, k-1)$ и суффикс $S(w, k-1)$ длины $k-1$ исходного слова $w = a_1 a_2 ... a_n$, а также указать условия, при которых решение возможно.

3. Метод определения префикса и суффикса

Предварительно отметим, что основная проблема (и в аспекте задачи реконструкции, и в аспекте задачи определения суффикса и префикса) заключается в том, что нам исходно дано мультимножество подслов V, а не кортеж подслов. При этом основная трудность связана именно с потерей порядка на исходных подсловах, полученных оператором сдвига один.

Решение поставленной задачи начнем с построения матрицы A, состоящей из m строк и k столбцов, строками которой являются исходные слова v_i из множества V. Слова из множества V представимы в виде $v_i = a_{i1}, a_{i2}, ... a_{ik}$, и элементами матрицы A являются символы алфавита $\sum - A = (a_{ij})$, где a_{ij} – символ алфавита на j-й позиции в i-м слове мультимножества V в порядке их перечисления.

Запишем явно матрицу A в прямой последовательности окна сдвига один. Очевидно, что в реальности в порядке перечисления по мультимножеству V мы будем наблюдать некоторую перестановку слов прямой последовательности, и, следовательно, соответственную перестановку строк матрицы A:

$$A = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_{m-1} \\ v_m \end{pmatrix} = \begin{pmatrix} a_1, a_2, ..., a_k \\ a_2, a_3, ..., a_{k+1} \\ a_3, a_4, ..., a_{k+2} \\ \vdots \\ a_{n-k}, a_{n-k+1}, ..., a_{n-1} \\ a_{n-k+1}, a_{n-k+2}, ..., a_n \end{pmatrix}.$$

Содержательно решение поставленной задачи опирается на анализ соседних столбцов этой матрицы. Рассмотрим первый и второй столбцы. В каждом из них при любой перестановке строк будет символ, стоящий на втором месте в неизвестном слове $w - a_2$, и символ, стоящий на третьем месте – a_3, и т.д. Если из этих двух столбцов вычеркнуть совпадающие пары символов, то останутся только символы a_1 и a_{n-k+2}. При условии, что они различны, мы получаем их конкретные значения. Если же a_1 и a_{n-k+2} совпадают, то будут вычеркнуты все символы в этих столбцах, и мы получаем информацию о том, что в соответствующих позициях префикса и суффикса находятся неизвестные, но совпадающие символы. Такой анализ может быть продолжен для всех $k-1$ пар соседних столбцов матрицы A. При условии, что после вычеркивания пар совпадающих символов у нас всегда остается не совпадающая пара, мы восстанавливаем префикс и суффикс длины $k-1$ неизвестного слова w.

Опишем метод формально.

Введем в рассмотрение кортеж всех символов алфавита, для которого разрешены кратности элементов

$$C = \left(s_1^{(\alpha 1)}, s_2^{(\alpha 2)}, ..., s_l^{(\alpha l)} \right),$$

при этом кратность 0 приводит к пустому множеству в данной позиции $s_i^{(0)} = \varnothing$. Определим оператор G действующий на i-й столбец матрицы A и создающий кортеж C_i, содержащий для всех символов алфавита их кратности в соответствии с числом символов, находящихся в этом столбце

$$GC(A, i) = C_i = \left(s_1^{(\alpha 1)}, s_2^{(\alpha 2)}, ..., s_l^{(\alpha l)} \right).$$

Применим оператор G к двум столбцам матрицы A, и обозначим:

$$GC(A, i) = C_i = \left(s_1^{(\alpha 1)}, s_2^{(\alpha 2)}, ..., s_l^{(\alpha l)} \right),$$
$$GC(A, k) = C_k = \left(s_1^{(\beta 1)}, s_2^{(\beta 2)}, ..., s_l^{(\beta l)} \right).$$

Введем в рассмотрение оператор получения символа GS, действующий на два кортежа столбцов матрицы A по следующему правилу:

$$GS(A,i,k) = \begin{cases} \bigcup_{j=1}^{l} s_j^{(\alpha_j - \beta_j)}, \left(s_j^{(\alpha_j)} \in GC(A,i) \right), \\ \qquad s_j^{(\beta_j)} \in GC(a,k), \\ s_j^{(\alpha j - \beta_j)} = \varnothing, \text{если } a_j - \beta_j \le 0. \end{cases}$$

Теперь применим оператор GS к двум последовательным столбцам матрицы A. В силу описанной выше структуры последовательных столбцов матрицы A результатом оператора GS будет

или символ, или пустое множество. Отметим, что если $GS(A, i, i+1) \neq \varnothing$, то и $GS(A, i+1, i) \neq \varnothing$. В этом случае мы определяем i-й символ префикса $a_i = GS(A, i, i+1)$ и $n-k+i$-й символ неизвестного слова $a_{n-k+i} = GS(A, i+1, i)$, который является i-м символом суффикса длины $k-1$.

Например, если $GS(A, 1, 2) = s_i$, то нам становится известен первый символ неизвестного слова w (первый символ префикса) – $a_1 = s_i$. В этой ситуации значение $GS(A, 2, 1)$ обязательно не пусто. Пусть $GS(A, 2, 1) = s_j$, в результате мы получаем первый символ суффикса $a_{n-k+2} = s_j$. Если же $GS(A, 1, 2) \neq \varnothing$, то очевидно, что и $GS(A, 2, 1) = \varnothing$, и мы получаем информацию о том, что $a_1 = a_{n-k+2}$. Однако при этом сам символ алфавита на этих позициях остается нам неизвестен.

Поскольку мы имеем $k-1$ последовательных пар столбцов, то если для каждой последовательной пары столбцов оператор GS возвращает непустое множество, то используя операцию «+» для обозначения конкатенации символов, мы получаем решение:

$$P(w, k-1) = a_1 a_2 \ldots a_{k-1} = \sum_{i=1}^{k-1} GS(A, i, i+1),$$

$$S(w, k-1) = a_{n-k+2} \ldots a_n = \sum_{i=1}^{k-1} GS(A, i+1, i).$$

Если для каждой пары оператор GS возвращает пустое множество, то символы префикса и суффикса остаются неизвестными, но при этом мы получаем информацию об их равенстве как подслов:

$$P(w, k-1) = S(w, k-1).$$

В общем случае мы получаем информацию о символах префикса и суффикса в виде некоторого шаблона, причем если это конкретные символы, то они расположены в одинаковых позициях префикса и суффикса, а если символы не удается определить, то у нас есть информация о том, что на этих позициях символы префикса и суффикса совпадают.

Приведем пример для слова $w = abbaaabb$ в алфавите $\sum = \{a, b\}$ и множества подслов, полученных оператором сдвига один с шириной окна, равной трем. При этом $k = 3$, $m = 6$, $n = 8$, и матрица A имеет вид:

$$A = \begin{pmatrix} abb \\ bba \\ baa \\ aaa \\ aab \\ abb \end{pmatrix}.$$

Применение оператора G к трем столбцам матрицы A дает следующие кортежи:

$$GC(A, 1) = C_1 = (a^{(4)}, b^{(2)}),$$

$$GC(A, 2) = C_2 = (a^{(3)}, b^{(3)}),$$

$$GC(A, 3) = C_3 = (a^{(3)}, b^{(3)}).$$

В результате мы получаем $GC(A, 1, 2) = a$, $GC(A, 2, 1) = b$, и $GC(A, 2, 3) = GC(A, 3, 2) = \varnothing$, и, тем самым, шаблоны префикса слова $w = abbaaabb$ длины два $P(w, 2) = a*$ и суффикса $S(w, 2) = b*$, где символ $*$ обозначает неизвестный, но совпадающий символ в соответствующих позициях префикса и суффикса (на самом деле это символ «b»).

4. Применение к задаче реконструкции

В одной из предыдущих статей [26] авторы предложили решение задачи о полной реконструкции в условиях мультимножества подслов и гипотезы сдвига один. В ряде случаев число реконструкций, определяемых числом эйлеровых путей или циклов в соответствующем мультиорграфе де Брейна, может быть значительным [26].

Введем в рассмотрение множество возможных реконструкций слов по исходному множеству V

$$W = \left\{(w| \ |w| = m, k-1 = n, V = SH1(w, k)\right\},$$

при этом если $|W| \geq 2$, то реконструкция возможна и многозначна. Пусть $w*$ – исходное, но неизвестное нам слово, по которому получено множество $V = SH1(w*, k)$. Тогда при выборе из возможных реконструкций (т.е. из множества W) мы выбираем только те слова, которые обладают полученным оператором GS префиксом и суффиксом, с учетом шаблонов неизвестных символов:

$$\tilde{W} = \left\{ \begin{array}{l} (w| \ P(w, k-1) = \sum_{i=1}^{k-1} GS(A, i, i+1), \\ S(w, k-1) = \sum_{i=1}^{k-1} GS(A, i+1, i) \end{array} \right\},$$

при этом гарантированно $w* \in \tilde{W}$.

Это приводит к редукции полученного множества реконструкций, поскольку мы рассматриваем только те слова, которые обладают заданными шаблонами префикса и суффикса. Более того, этот подход можно применять не только для редукции

конечного множества реконструкций, а рассматривать префикс как шаблон выбора начальных дуг для эйлеровых путей в мультиорграфе де Брейна при построении реконструкции [26].

Заключение

В статье в аспекте решения задачи восстановления символьных описаний временных рядов и логов бизнес-процессов предложено решение задачи определения префикса и суффикса неизвестного слова. Решение основано на предположении о том, что исходно задано полное множество подслов фиксированной длины k, порожденное смещением окна длины k по неизвестному слову со сдвигом один. Получено решение, позволяющее получить информацию о префиксе и суффиксе неизвестного слова или некоторый шаблон для префикса и суффикса. Предложенное решение позволяет получить дополнительную информацию о возможных реконструкциях и тем самым сократить число возможных реконструкций слов по заданному множеству подслов. В лучшем случае предложенный метод позволяет определить префикс и суффикс длины k неизвестного слова, а в худшем случае – констатировать, что префикс и суффикс совпадают между собой.

Результаты могут быть использованы совместно с решением задачи реконструкции [26, 27] для редукции множества возможных реконструкций при качественном анализе в таких задачах бизнес-информатики, как анализ временных рядов и логов бизнес-процессов. ■

Благодарности

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-07-00150.

Литература

1. Querying and mining of time series data: Experimental comparison of representations and distance measures / H. Ding [et al.] // Proceedings of the VLDB Endowment. 2008. Vol. 1. No 2. P. 1542–1552. DOI: 10.14778/1454159.1454226.
2. Kurbalija V., Radovanović M., Geler Z., Ivanović M. The influence of global constraints on DTW and LCS similarity measures for time-series databases // Advances in Intelligent and Soft Computing. 2011. Vol. 101. P. 67–74. DOI: 10.1007/978-3-642-23163-6_10.
3. Wu Y.-L., Agrawal D., el Abbadi A. A comparison of DFT and DWT based similarity search in time-series databases // Ninth International Conference on Information and Knowledge Management (CIKM '00), McLean, VA, 6–11 November 2000. P. 488–495.
4. Bemdt D.J., Clifford J. Using dynamic time warping to find patterns in time series // AAAI-94 Workshop on Knowledge Discovery in Databases. 1994. P. 359–370. [Электронный ресурс]: https://www.aaai.org/Papers/Workshops/1994/WS-94-03/WS94-03-031.pdf (дата обращения 15.03.2020).
5. Dreyer W., Dittrich A.K., Schmidt D. Research perspectives for time series management systems // SIGMOD Record. 1994. Vol. 23. No 1. P. 10–15.
6. Keogh E.J., Pazzani M.J. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback // Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98), New York, 27–31 August 1998. P. 239–241.
7. Andersen B. Business processes improvement toolbox. New York: ASQ Quality Press, 1999.
8. Lin J., Keogh E., Wei L., Lonardi S. Experiencing SAX: A novel symbolic representation of time series // Data Mining and Knowledge Discovery. 2007. Vol. 15. No 2. P. 107–144. DOI: 10.1007/s10618-007-0064-z.
9. String reconstruction from substring compositions / J. Acharya [et al.] // SIAM Journal on Discrete Mathematics. 2014. Vol. 29. No 3. P. 1340–1371.
10. Reconstruction of sequences / B. Manvel [et al.] // Discrete Mathematics. 1991. Vol. 94. No 3. P. 209–219. DOI: 10.1016/0012-365X(91)90026-X.
11. Carpi A., de Luca A. Words and special factors // Theoretical Computer Science. 2001. Vol. 259. No 1–2. P. 145–182.
12. de Luca A. On the combinatorics of finite words // Theoretical Computer Science. 1999. Vol. 218. No 1. P. 13–39.
13. Dudík M., Schulman L.J. Reconstruction from subsequences // Journal of Combinatorial Theory. Series A. 2003. Vol. 103. No 2. P. 337–348. DOI: 10.1016/S0097-3165(03)00103-1.
14. Erdös P.L., Ligeti P., Sziklai P., Torney D.C. Subwords in reverse-complement order // Annals of Combinatorics. 2006. Vol. 10. No 4. P. 415–430. DOI: 10.1007/s00026-006-0297-3.
15. Fici G., Mignosi F., Restivo A., Sciortino M. Word assembly through minimal forbidden words // Theoretical Computer Science. 2006. Vol. 359. No 1–3. P. 214–230. DOI: 10.1016/j.tcs.2006.03.006.
16. Levenshtein V.I. Efficient reconstruction of sequences from their subsequences or supersequences // Journal of Combinatorial Theory, Series A. 2001. Vol. 93. P. 310–332.
17. Piña C., Uzcátegui C. Reconstruction of a word from a multiset of its factors // Theoretical Computer Science. 2008. Vol. 400. No 1–3. P. 70–83. DOI: 10.1016/j.tcs.2008.01.052.

18. Lothaire M. Algebraic combinatorics on words. Cambridge, UK: Cambridge University Press, 2002.

19. Gusfield D. Algorithms on strings, trees, and sequences: Computer science and computational biology. Cambridge, UK: Cambridge University Press, 1997.

20. Skiena S.S., Sundaram G. Reconstructing strings from substrings // Journal of Computational Biology. 1995. Vol. 2. No 2. P. 333–353.

21. Leont'ev V.K., Smetanin Y.G. Problems of Information on the set of words // Journal of Mathematical Sciences. 2002. Vol. 108. No 1. P. 49–70. DOI: 10.1023/A:1012705332306.

22. Левенштейн В.И. Восстановление объектов по минимальному числу искаженных образцов // Доклады РАН. 1997. Т. 354. № 5. С. 593–596.

23. Krasikov I., Roditty Y. Note: On a reconstruction problem for sequences // Journal of Combinatorial Theory. Series A. 1997. No 77. P. 344–348.

24. Ульянов М.В., Сметанин Ю.Г. Подход к определению характеристик колмогоровской сложности временных рядов на основе символьных описаний // Бизнес-информатика. 2013. № 2. С. 49–54.

25. Сметанин Ю.Г., Ульянов М.В. Мера символьного разнообразия: подход комбинаторики слов к определению обобщенных характеристик временных рядов // Бизнес-информатика. 2014. № 3. С. 40–46.

26. 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. 2014. Vol. 50. No 1. P. 148–156.

27. 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. 2015. Vol. 51. No 1. P. 157–164. DOI: 10.1007/s10559-015-9708-y.

Об авторах

Жукова Галина Николаевна

кандидат физико-математических наук;

доцент департамента программной инженерии, факультет компьютерных наук,
Национальный исследовательский университет «Высшая школа экономики»,
101000, г. Москва, ул. Мясницкая, д. 20;

E-mail: galinanzhukova@gmail.com

ORCID: 0000-0003-1835-7422

Сметанин Юрий Геннадиевич

доктор физико-математических наук:

главный научный сотрудник, Федеральный исследовательский центр «Информатика и управление» Российской академии наук,
119333, г. Москва, ул. Вавилова, д. 40;

E-mail: smetanin.iury2011@yandex.ru

ORCID: 0000-0003-0242-6972

Ульянов Михаил Васильевич

доктор технических наук, профессор;

ведущий научный сотрудник, Институт проблем управления им. В.А. Трапезникова Российской академии наук,
117997, г. Москва, ул. Профсоюзная, д. 65;

E-mail: muljanov@mail.ru

ORCID: 0000-0002-5784-9836

About the possibility of determining the prefix and suffix of a word by subwords of fixed length

Galina N. Zhukova[a]
E-mail: galinanzhukova@gmail.com

Yuri G. Smetanin[b]
E-mail: smetanin.iury2011@yandex.ru

Mikhail V. Ulyanov[c]
E-mail: muljanov@mail.ru

[a] National Research University Higher School of Economics
Address: 20, Myasnitskaya Street, Moscow 101000, Russia
[b] Federal Research Center "Computer Science and Control", Russian Academy of Sciences
Address: 40, Vavilova Street, Moscow 119333, Russia
[c] Trapeznikov Institute of Control Sciences, Russian Academy of Sciences
Address: 65, Profsoyuznaya Street, Moscow 117997, Russia

Abstract

In applied problems of business informatics related to data analysis (in particular, in the analysis and forecasting of time series, in the study of log files of business processes, etc.), problems of qualitative analysis arise. Qualitative analysis methods often use symbolic coding as a way of presenting information about the processes under study. In a number of situations, due to the fragmentation of such descriptions, the problem arises of reconstructing a complete symbolic description of a process (word) from its successive fragments (subwords). From the multiset of all subwords of a sufficiently large length, the original word is uniquely restored. In the case of insufficiently long subwords, several different reconstructions of the original word are possible. The number of feasible reconstructions can be reduced by determining the suffix and prefix of the reconstructed word. A method is proposed for determining the prefix and suffix of a word consisting of $k-1$ symbols each on the basis of multiset V of subwords of a fixed length equal to k. We accept the hypothesis that this multiset is generated by a window of a fixed length k of one symbol shift in an unknown word. The method for determining the prefix and suffix is based on the construction and analysis of the matrix formed by subwords from V written in rows in arbitrary order and the use of the operator acting on multisets of characters of the alphabet formed by neighboring columns of this matrix. The method is capable of determining the prefix $a_1a_2 \ldots a_{k-1}$ and suffix $b_1b_2 \ldots b_{k-1}$, if $a_i \neq b_i$ for any i from 1 to $k-1$. If in the prefix and suffix $a_i \neq b_i$ only for some values of i, the characters in the corresponding positions are determined, and $a_j = b_j$ for the remaining characters. In the worst case, the method concludes that $a_i = b_i$ for any i from 1 to $k-1$, but does not determine the characters themselves. This is a situation in which the prefix and suffix coincide but cannot be determined.

Key words: word reconstruction; prefix; suffix; multiset of subwords; subwords of fixed length; shift operator.

Citation: Zhukova G.N., Smetanin Yu.G., Ulyanov M.Yu. (2020) About the possibility of determining the prefix and suffix of a word by subwords of fixed length. *Business Informatics*, vol. 14, no 2, pp. 84–92.
DOI: 10.17323/2587-814X.2020.2.84.92

References

1. Ding H., Trajcevski G., Scheuermann P., Wang X., Keogh E. (2008) Querying and mining of time series data: Experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment*, vol. 1, no 2, pp. 1542–1552. DOI: 10.14778/1454159.1454226.
2. Kurbalija V., Radovanović M., Geler Z., Ivanović M. (2011) The influence of global constraints on DTW and LCS similarity measures for time-series databases. *Advances in Intelligent and Soft Computing*, vol. 101, pp. 67–74. DOI: 10.1007/978-3-642-23163-6_10.
3. Wu Y.-L., Agrawal D., el Abbadi A. (2000) A comparison of DFT and DWT based similarity search in time-series databases. Proceedings of the *Ninth International Conference on Information and Knowledge Management (CIKM '00), McLean, VA, 6–11 November 2000*, pp. 488–495.
4. Bemdt D.J., Clifford J. (1994) Using dynamic time warping to find patterns in time series. *AAAI-94 Workshop on Knowledge Discovery in Databases*, pp. 359–370. Available at: https://www.aaai.org/Papers/Workshops/1994/WS-94-03/WS94-03-031.pdf (accessed 15 March 2020).
5. Dreyer W., Dittrich A.K., Schmidt D. (1994) Research perspectives for time series management systems. *SIGMOD Record*, vol. 23, no 1, pp. 10–15.
6. Keogh E.J., Pazzani M.J. (1998) An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. Proceedings of the *Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98), New York, 27–31 August 1998*, pp. 239–241.
7. Andersen B. (1999) *Business processes improvement toolbox*. New York: ASQ Quality Press.
8. Lin J., Keogh E., Wei L., Lonardi S. (2007) Experiencing SAX: A novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, vol. 15, no 2, pp. 107–144. DOI: 10.1007/s10618-007-0064-z.
9. Acharya J., Das H., Milenkovic O., Orlitsky A., Pan S. (2014) String reconstruction from substring compositions. *SIAM Journal on Discrete Mathematics*, vol. 29, no 3, pp. 1340–1371.
10. Manvel B., Meyerowitz A., Schwenk A., Smith K., Stockmeyer P. (1991) Reconstruction of sequences. *Discrete Mathematics*, vol. 94, no 3, pp. 209–219. DOI: 10.1016/0012-365X(91)90026-X.
11. Carpi A., de Luca A. (2001) Words and special factors. *Theoretical Computer Science*, vol. 259, no 1–2, pp. 145–182.
12. de Luca A. (1999) On the combinatorics of finite words. *Theoretical Computer Science*, vol. 218, no 1, pp. 13–39.

13. Dudík M., Schulman L.J. (2003) Reconstruction from subsequences. *Journal of Combinatorial Theory. Series A*, vol. 103, no 2, pp. 337–348. DOI: 10.1016/S0097-3165(03)00103-1.

14. Erdös P.L., Ligeti P., Sziklai P., Torney D.C. (2006) Subwords in reverse-complement order. *Annals of Combinatorics*, vol. 10, no 4, pp. 415–430. DOI: 10.1007/s00026-006-0297-3.

15. Fici G., Mignosi F., Restivo A., Sciortino M. (2006) Word assembly through minimal forbidden words. *Theoretical Computer Science*, vol. 359, no 1–3, pp. 214–230. DOI: 10.1016/j.tcs.2006.03.006.

16. Levenshtein V.I. (2001) Efficient reconstruction of sequences from their subsequences or supersequences. *Journal of Combinatorial Theory, Series* A, Vol. 93, pp. 310–332.

17. Piña C., Uzcátegui C. (2008) Reconstruction of a word from a multiset of its factors. *Theoretical Computer Science*, vol. 400, no 1–3, pp. 70–83. DOI: 10.1016/j.tcs.2008.01.052.

18. Lothaire M. (2002) *Algebraic combinatorics on words*. Cambridge, UK: Cambridge University Press.

19. Gusfield D. (1997) *Algorithms on strings, trees, and sequences: Computer science and computational biology*. Cambridge, UK: Cambridge University Press.

20. Skiena S.S., Sundaram G. (1995) Reconstructing strings from substrings. *Journal of Computational Biology*, vol. 2, no 2, pp. 333–353.

21. Leont'ev V.K., Smetanin Y.G. (2002) Problems of Information on the set of words. *Journal of Mathematical Sciences*, vol. 108, no 1, pp. 49–70. DOI: 10.1023/A:1012705332306.

22. Levenshtein V.I. (1997) Restoring objects based on the minimum number of distorted samples. *Doklady Akademii Nauk*, vol. 354, no 5, pp. 593–596 (in Russian).

23. Krasikov I., Roditty Y. (1997) Note: On a reconstruction problem for sequences. *Journal of Combinatorial Theory, Series A*, no 77, pp. 344–348.

24. Ulyanov M.V., Smetanin Yu.G. (2013) Determining the characteristics of Kolmogorov complexity of time series: An approach based on symbolic descriptions. *Business Informatics*, no 2, pp. 49–54 (in Russian).

25. Smetanin Yu.G., Ulyanov M.V. (2014) Measure of symbolical diversity: Combinatorics on words as an approach to identify generalized characteristics of time series. *Business Informatics*, no 3, pp. 40–46 (in Russian).

26. Smetanin Yu.G., Ulyanov M.V. (2014) 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*, vol. 50, no 1, pp. 148–156.

27. Smetanin Yu.G., Ulyanov M.V. (2015) 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*, vol. 51, no 1, pp. 157–164. DOI: 10.1007/s10559-015-9708-y.

About the authors

Galina N. Zhukova

Cand. Sci. (Phys.-Math.);

Associate Professor, School of Software Engineering, Faculty of Computer Science,
National Research University Higher School of Economics,
20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: galinanzhukova@gmail.com

ORCID: 0000-0003-1835-7422

Yuri G. Smetanin

Dr. Sci. (Phys.-Math.);

Chief Researcher, Federal Research Center "Computer Science and Control", Russian Academy of Sciences,
40, Vavilova Street, Moscow 119333, Russia;

E-mail: smetanin.iury2011@yandex.ru

ORCID: 0000-0003-0242-6972

Mikhail V. Ulyanov

Dr. Sci. (Tech.);

Leading Researcher, Laboratory of Scheduling Theory and Discrete Optimization, V.A. Trapeznikov Institute of Control Sciences,
Russian Academy of Sciences,
65, Profsoyuznaya Street, Moscow 117997, Russia;

E-mail: muljanov@mail.ru

ORCID: 0000-0002-5784-9836