УДК: 519.24
DOI: 10.25559/SITITO.16.202001.23-32

Метод определения периода зашумленной периодической символьной последовательности, основанный на позициях подслов в последовательности

Г. Н. Жукова[1], А. В. Жуков[2], Ю. Г. Сметанин[3], М. В. Ульянов[4,5*]

[1] Национальный исследовательский университет «Высшая школа экономики», г. Москва, Россия
101000, Россия, г. Москва, ул. Мясницкая, д. 20

[2] Национальный исследовательский центр «Курчатовский институт», г. Москва, Россия
123182, Россия, г. Москва, пл. Академика Курчатова, д. 1

[3] Федеральный исследовательский центр «Информатика и управление» Российской академии наук, г. Москва, Россия
119333, Россия, г. Москва, ул. Вавилова, д. 44, корп. 2

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

[5] Московский государственный университет имени М.В. Ломоносова, г. Москва, Россия
119991, Россия, г. Москва, ГСП-1, Ленинские горы, д. 1

* muljanov@mail.ru

Аннотация

Предложен метод определения периода искаженной шумом периодической последовательности. Период почти периодической последовательности — это длина наименьшего периодически повторяющегося фрагмента, образующего соответствующую обычную периодическую последовательность. Метод может быть применен для искаженных периодических последовательностей, полученных из периодических последовательностей, состоящих из, по крайней мере, восьми полных периодически повторяющихся минимальных фрагментов.

В соответствующих периодических последовательностях с шумом замены, вставки и удаления некоторые периодические фрагменты могут быть искажены из-за внесения шума. Уровень шума предполагается менее 10%, это предположение позволяет использовать оператор сдвига с окном ширины 16 и наблюдать в этом окне более двух раз каждый неповрежденный фрагмент длины 16, содержащийся в исследуемой последовательности.

Метод основан на подсчете числа символов в слове w между первыми символами ближайших одинаковых подслов длины 16. Для вычисления разностей между левыми позициями соседних одинаковых подслов используются только подслова, встретившиеся в рассматриваемом слове w более двух раз. Все найденные разности располагаются в порядке возрастания и находятся квантиль 25% и медиана в последовательности разностей.

Вычислительный эксперимент показал, что 25% квантиль дает удовлетворительную оценку периода при уровне шума менее 5 %. Иногда метод дает достаточно хороший результат в случае шума от 5 до 10 %. Зависимость доли удовлетворительных оценок периода от уровня шума исследовалась для каждого типа шума отдельно, а также для смеси шумов всех трех типов в одинаковых пропорциях. Вычислительный эксперимент показал, что 25% квантиль дает более точную оценку периода, чем медиана. Предполагается улучшить метод таким образом, чтобы восстанавливать саму периодическую последовательность только по последовательности с шумом.

Ключевые слова: символьная последовательность, периодическая последовательность, последовательность с шумами, шум вставки, шум удаления, шум замены.

Финансирование: исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований в рамках научного проекта № 19-07-00151 «Комплексные научные и экспериментальные междисциплинарные исследования по моделированию поведения разливов нефти на море под воздействием гидрометеорологических условий».

Для цитирования: Жукова, Г. Н. Метод определения периода зашумленной периодической символьной последовательности, основанный на позициях подслов в последовательности / Г. Н. Жукова, А. В. Жуков, Ю. Г. Сметанин, М. В. Ульянов. – DOI 10.25559/SITITO.16.202001.23-32 // Современные информационные технологии и ИТ-образование. – 2020. – Т. 16, № 1. – С. 23-32.

The Method of Estimating the Period of a Symbolic Periodic Sequence with Noise, Based on the Sub-Words Positions in the Sequence

G. N. Zhukova[a], A. V. Zhukov[b], Yu. G. Smetanin[c], M. V. Ulyanov[d,e*]
[a] National Research University Higher School of Economics, Moscow, Russia
20 Myasnitskaya St., Moscow 101000, Russia
[b] National Research Center "Kurchatov Institute", Moscow, Russia
1 Akademika Kurchatova pl., Moscow 123182, Russia
[c] Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia. 44-2 Vavilova St., Moscow 119333, Russia
[d] V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russia
65 Profsoyuznaya St., Moscow 117997, Russia
[e] Lomonosov Moscow State University, Moscow, Russia. 1, Leninskie gory, Moscow 119991, Russia
[*] muljanov@mail.ru

Abstract

The method of estimating the period of an almost periodic sequence is proposed. The period of an almost periodic sequence is the length of the smallest periodic pattern that composes the corresponding completely periodic sequence by repetition. The method can be applied to corrupted periodic sequences that originated from a periodic sequence that consists of no less than eight complete periodic patterns. In the corresponding periodic sequence with noise of insertion, deletion or change, some periodic patterns might be corrupted by noise intrusion. The level of noise is supposed to be less than 10%, this assumption allows us to use a shift operator with the window of width 16 so that we can observe at least more than twice every uncorrupted pattern of length 16 that the sequence under consideration contains. The method is based on the counting of the number of symbols in the word w between the two first symbols of the closest equal patterns of length 16. Only patterns that were encountered in word w more than twice are used for counting the differences between the left positions of the neighbor equal patterns in the sequence under consideration. All the differences counted are arranged in the ascending order and the 25% quantile and the median of the sequence of the differences are calculated. The computational experiment shows that in most cases the 25% quantile is an appropriate estimation of the length of the periodic pattern when the noise level is no more than 5%. Sometimes the method provides a sufficient result in case of noise level between 5 and 10%. The dependence of the percentage of sufficient estimations of periodic pattern length on the noise level was studied in cases of only one type of noise and the noise of all three types in equal proportions. The computational experiment showed that in all situations 25% quantile provides more sufficient estimations than the median. The method is supposed to be improved, in order to restore the whole corresponding periodic sequence using only the sequence with noise.

Keywords: symbolic sequence, periodic sequence, sequence with noise, noise of insertion, noise of deletion, noise of change.

Funding: This work was prepared with the financial support of the Russian Foundation for Basic Research in the framework of scientific project No. 19-07-00151 "Complex Scientific and Experimental Interdisciplinary Studies of Modeling the Behavior of Oil Spills at Sea Under The Influence of Hydrometeorological Conditions".

For citation: Zhukova G.N., Zhukov A.V., Smetanin Yu.G., Ulyanov M.V. The Method of Estimating the Period of a Symbolic Periodic Sequence with Noise, Based on the Sub-Words Positions in the Sequence. *Sovremennye informacionnye tehnologii i IT-obrazovanie* = Modern Information Technologies and IT-Education. 2020; 16(1):23-32. DOI: https://doi.org/10.25559/SITITO.16.202001.23-32

1. Введение

Среди различных задач, связанных с анализом больших данных, можно выделить группу задач качественного анализа. Такой анализ актуален для различных прикладных областей, в которых исследователей интересуют особенности качественного поведения исследуемых объектов или процессов. Для решения задач качественного анализа может быть эффективно использован метод символьного кодирования наблюдаемых состояний процесса с последующим анализом полученного символьного представления. Применение кодирования символами из конечного алфавита позволяет отбросить не существенные детали, которые не несут полезной информации в аспекте задач качественного анализа [1,2].

Отметим, что в настоящее время значительная часть исследуемых процессов обладает циклическим характером, и, при этом, достаточно часто исходные данные подвержены воздействию случайных искажений, вызванных внешними факторами. Укажем, например, такие процессы и связанные с ними задачи анализа, в том числе и задачи поиска периодов, как прогнозирование погоды, прогнозирование котировок на фондовых биржах, анализ различных особенностей клинических заболеваний, анализ кардиограмм, исследования в биоинформатике, например, поиск скрытых периодов в ДНК, а также задачи, связанные с анализом и прогнозом временных рядов [3-14]. При символьном кодировании циклический характер процессов отображается периодическими символьными последовательностями, а случайные искажения приводят к моделям периодической символьной последовательности с зашумленными циклами.

Очевидно, возникающая при анализе таких последовательностей задача нахождения длины периода осложняется при добавлении в периодическую последовательность различных видов шумов, таких как вставка, замена и удаление символов. Наличие таких шумов не позволяет использовать классические алгоритмы поиска периода в символьных последовательностях [15]. При этом шумы вставки и удаления искажают не только последовательность символов, составляющих период, но и его длину, что вносит дополнительные сложности в анализ. При шуме замены мы имеем некоторый шаблон с длиной равной периоду, каждая реализация которого может содержать искажения некоторых символов с некоторыми вероятностями. Для решения задачи поиска периодичности в зашумленных данных в настоящее время существует ряд алгоритмов и методов, среди которых наиболее известны: алгоритм WARP [16], алгоритм CONV [17], алгоритм STNR [18, 19] и метод, предложенный коллективом авторов под руководством Е.В. Короткова [20, 21, 22]. Проведенный нами ранее анализ данных алгоритмов [23] показал, что они обладают различной чувствительностью по отношению к шумам различных типов. Интерес представляет разработка такого метода идентификации длины периода в зашумленных символьных последовательностях, который будет инвариантен к типу шума. В качестве основы такого метода авторы предлагают использовать аппарат вычисления энтропии конечных слов, базирующийся на определении встречаемости подслов фиксированной длины в исследуемой символьной последовательности.

2. Терминология и обозначения

Мы будем использовать далее следующую терминологию и обозначения, введенную ранее в [24]:

$\Sigma = \{s_1, s_2, \ldots, s_m\}$ — конечный алфавит, s — произвольный символ алфавита;

Σ^k — k-ая декартова степень множества Σ, т.е. множество всех кортежей длины k;

$a \in \Sigma^k$ — k-элементный кортеж, $a = (a_1, a_2, \ldots, a_k)$, $a_i \in \Sigma$;

w — слово (над алфавитом Σ) — последовательность символов алфавита;

$|w| = n$ — длина слова, понимаемая как число символов слова;

$SW(w, i, l)$ — оператор, выделяющий подслово длины l в слове w, начиная с символа в позиции i. Пусть $|w| = n$, тогда оператор определен при $i + l - 1 \le n$: $SW(s_1 s_2 \ldots s_n, i, l) = s_i s_{i+1} \ldots s_{i+l-1}$;

$SH1(w, l)$ — оператор сдвига 1, основанный на операторе $SW(w, i, l)$, и действующий на слово w окном ширины l. Определенный при $|w| \ge l$ оператор порождает множество из k уникальных подслов длины l выполняя сдвиг на единицу окна ширины l по слову w, начиная с крайней левой позиции слова w.

3. Особенности частотной встречаемости слов в символьных последовательностях

Воздействуя оператором сдвига один $SH1(w, l)$ с окном ширины l на исследуемое слово w, $|w| = n$, мы получим множество из k уникальных подслов $v_i, i = \overline{1,k}$ длины l. Каждому из этих k уникальных подслов сопоставим последовательность номеров $i_1, \ldots, i_r$ символов в слове w, таких, что последовательные l символов слова w, начиная с символа с номером $i_j, j = \overline{1,r}$ совпадают с соответствующим подсловом v_i. Обозначим $c_i, i = \overline{1,k}$ количество номеров, соответствующую подслову v_i.

В дальнейшем мы рассматриваем слова над алфавитом $\Sigma = \{0,1\}$. Рассмотрим, как особенности исследуемого слова w влияют количество уникальных подслов k, и на значения $c_i, i = \overline{1,k}$.

1. *Случайное слово.* Пусть слово w получено генератором псевдослучайных чисел с равномерным распределением (символы 0 и 1 равновероятны). При $l = 1$ существует не более двух уникальных подслов, а именно 0 и 1. В этом случае значения $c_1 \oplus c_2$, поскольку символы бинарного алфавита имеют слабо отличающуюся частотную встречаемость в случайных словах большой длины. При $l = 2$ мы можем наблюдать четыре возможных подслова — 00, 01, 10, 11, и соответствующие значения $c_i, i = \overline{1,4}$ также близки при больших n. Рассмотрим значения c_i для уникальных подслов при условии, что мы видим все возможные подслова длины l — $c_i, i = \overline{1,k}$, где $k = 2^l$. Если все значения c_i мало отличаются друг от друга, то это характеризует исследуемое слово w, как слово, обладающее полным равночастотным разнообразием в области подслов малой длины $l : n \ge 2^l$, что характеризует его случайность.

2. *Периодическое слово.* Пусть слово w является периодическим с периодом p, который многократно повторяется до достижения длины n. Допустим, что подслово u длины p, задающее период, также получено генератором псевдослучайных чисел с равномерным распределением. При этом мы также наблюдаем очень близкие значения c_i, однако не наблюдаем

полного разнообразия подслов, поскольку мы видим не более p подслов из 2^l возможных. Если минимальный периодически повторяющийся фрагмент не содержит дважды некоторое подслово (т.е. $s_{i_1},\ldots,s_{i_j}=s_{i_{1+t}},\ldots,s_{i_{j+t}}$ только при $t=kp$, k — целое), то все уникальные подслова встречаются одинаково часто (только последние $l-1$ подслов встретятся на один раз меньше в последовательности, состоящей из целого числа периодически повторяющихся фрагментов). Таким образом, оценка p^* длины периода равна k — числу наблюдаемых уникальных подслов. В предположении о случайном источнике подслова, образующего минимальный периодически повторяющийся фрагмент исследуемой символьной последовательности, эта оценка будет равна истинной длине периода при достаточно большой ширине окна l в операторе сдвига один, поскольку при этом $2^l >> p$, что позволяет обеспечить уникальность подслов длины l при сдвиге один в рамках периода.

3. *Периодическое слово с шумом*. Оценим характер поведения c_i для периодического слова, в которое был внесен небольшой шум замены, удаления или вставки. Пусть период исходной последовательности равен p, минимальный периодически повторяющийся фрагмент u длины p получен генератором псевдослучайных чисел с равномерным распределением, слово длины n содержит целое число таких фрагментов. В слово w вносится случайный шум замены, удаления и вставки. Отметим, что для окна $l >> \lfloor \log_2 n \rfloor$ разнообразие различных подслов существенно больше, чем возможное число позиций окна. Заметим, что для такого окна внесение единичного шума замены, вставки или удаления в слово w приведет, скорее всего, к появлению l новых подслов длины l в операторе сдвига один. Эти слова встретятся ровно один раз, а уникальные подслова неискаженной последовательности встретятся в зашумленной последовательности на один раз меньше, чем в исходной. При достаточно большом числе полных минимальных фрагментов в исходной периодической последовательности (далее рассматриваем случай не менее восьми повторений) p уникальных подслов будут встречаться в последовательности примерно одинаково часто, причем более 2 раз, а уникальные слова, образовавшиеся из-за шума, встретятся по одному разу. Если уровень шума не слишком мал, то из-за шума возможно появление одного и того же нового подслова дважды, поэтому далее мы будем исключать из рассмотрения подслова, встретившиеся не более двух раз. Таким образом, детальный анализ встречаемости уникальных подслов позволяет сформулировать следующий метод идентификации длины периода.

4. Описание метода

На основе приведенных выше рассуждений авторы предлагают следующий метод идентификации длины периода в зашумленных символьных последовательностях:

1. Исследуемое слово w, $|w|=n$ обрабатывается оператором сдвига один $SH1(w,l)$ с окном ширины l. Рекомендуемое значение $l >> \lfloor \log_2 n \rfloor$. В результате получаем множество V из k уникальных подслов $v_i, i=\overline{1,k}$ длины l.

2. Каждому подслову v_i из V ставится в соответствие (возрастающая) последовательность номеров $i_1,\ldots,i_r$ символов в слове w, таких, что последовательные l символов слова w, начиная с символа с номером i_j, $j=\overline{1,r}$ совпадают с соответствующим подсловом v_i.

3. Подслова, которые встретились не более двух раз, исключаются из множества V, поскольку мы считаем, что их появление связано с шумом, а, значит, они не входили в исходную периодическую последовательность, период которой мы хотим найти.

4. По каждой последовательности номеров $i_1,\ldots,i_r$ строится множество разностей последовательных номеров. Разность последовательных номеров показывает, через сколько символов в исходном слове w встретилось подслово v_i после того, как оно встретилось предыдущий раз. Для периодической последовательности без искажений разности будут равны периоду для слов, которые встречаются в периодически повторяющемся фрагменте один раз. Если подслово встречается в периодически повторяющемся фрагменте длиной период более одного раза, то эта разность будет меньше периода.

5. Все множества разностей объединяются, их элементы упорядочиваются по возрастанию. В полученной последовательности находятся квартиль q_{25} и медиана. Если различие значений q_{25} и медианы не более 10% ($med/q_{25}<1.1$), то значение q_{25} считаем оценкой длины периода. Из-за шума подслово v_i в слове w не всегда будет повторяться ровно через период, но оно будет повторяться через промежутки, примерно равные целому числу периодов. Поэтому мы будем считать, что при не очень большом уровне шума повторения подслова v_i будут происходить через один или два периода. В последнем случае, если $1.9<med/q_{25}<2.1$, т.е. если с точностью порядка 10% медиана в два раза больше q_{25}, то значение q_{25} тоже считаем оценкой длины периода. Во всех остальных случаях считаем, что метод не нашел скрытой периодичности.

В рамках дальнейшего исследования интерес представляет изучение порога применимости метода, т.е. уровня шума, при котором данный подход еще может идентифицировать период. Такие данные могут быть получены экспериментально.

5. Описание вычислительного эксперимента

Для получения синтетических периодических символьных последовательностей с шумами был использован подход, основанный на вероятностной модели шумов, ранее предложенный авторами [25]. Мы кратко приводим описание основных этапов вычислительного эксперимента и способ создания последовательности со смешанным шумом.

1. С помощью генератора псевдослучайных чисел с равномерным распределением получаем последовательность u символов длины p, число p далее считаем истинным значением периода, слово u — периодически повторяющимся фрагментом. Повторяя периодический фрагмент k раз, получаем периодическую последовательность.

2. По полученной периодической последовательности строится последовательность с искажениями. Искажения вносятся в следующем порядке. Вначале задается уровень шума каждого вида — замены, вставки и удаления γ_{ch}, γ_{ins}, γ_{del}, не более 10%. По уровню шума вычисляется количество искаженных символов $n_t=\lceil \gamma_t \cdot p \cdot k \rceil$, $t=ch,ins,del$

а) С помощью генератора псевдослучайных чисел получаем позиции, в которых нужно удалить символ. Номера позиций получаем равномерным перемешиванием $p \cdot k$ последовательных натуральных чисел от 1 до $p \cdot k$, берем первые n_{del} символов перемешанной последовательности. Удаляем из исходной по-

следовательности символы с полученными номерами.

b) Затем вносим шум замены, для чего аналогично позициям удаления находятся позиции для замены, только на этот раз используется последовательность, полученная после удаления символов, так что перемешивается уже не $p \cdot k$, а $p \cdot k - n_{del}$ натуральных чисел. Из перемешанных $p \cdot k - n_{del}$ натуральных чисел берем первые n_{ch} чисел и инвертируем символы последовательности, полученной после внесения шума удаления.

c) После внесения шума замены добавляем символы, для этого перемешиваем $p \cdot k - n_{del}$ натуральных чисел и берем первые n_{ins} чисел. В последовательности, полученной на шаге b), после элементов с номерами, равными выбранным числам, вставляем ноль или единицу. Какое именно число вставляется, определяется генератором псевдослучайных чисел с равномерным распределением.

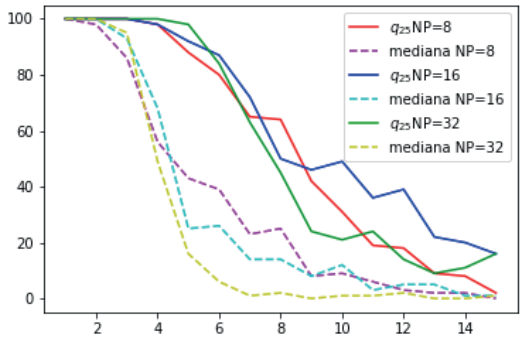
3. Для последовательности с шумом строим множество подслов длины 16, встречающихся в этой последовательности, для каждого подслова составляем последовательность номеров, начиная с которых в последовательности с шумом находится данное подслово.

4. На основе множества разностей номеров находится оценка периода в соответствии с описанным выше методом.

При проведении вычислительного эксперимента строились периодические последовательности с периодом от 20 до 2000, периодический фрагмент повторялся от 8 до 32 раз. Для каждого фиксированного уровня шума от 1% до 15% было построено 100 искаженных последовательностей, для каждой из них была вычислена оценка периода.

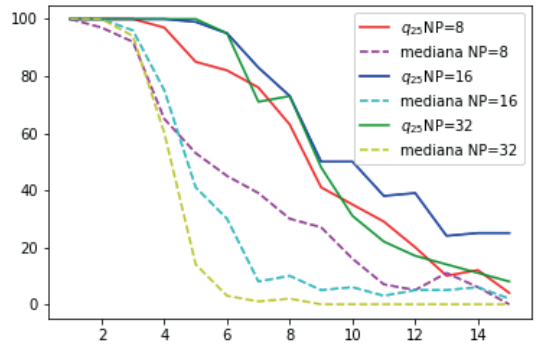
6. Экспериментальные результаты

Полученные экспериментальные результаты представлены ниже в виде графиков. На рис. 1 изображен график зависимости доли правильно распознанных периодов (в процентах) от уровня (тоже в процентах) шума замены. Правильно распознанным периодом считался случай, когда оценка периода по описанному выше методу отличалась от исходного периода не более чем на 10%. На рис. 2 и 3 представлены аналогичные зависимости соответственно для шума вставки и удаления. На рис. 4 представлена зависимость доли правильно распознанных периодов для смешанного шума, включающего шум всех трех типов в равных долях.



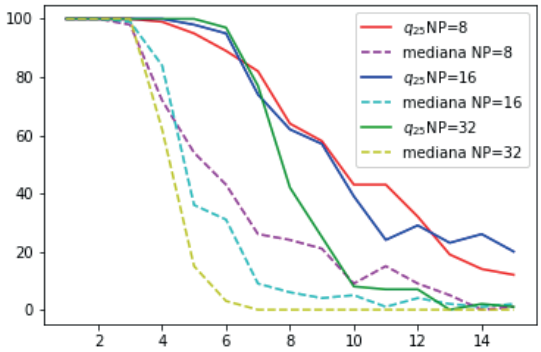
Р и с. 1. График зависимости доли правильно распознанных периодов от уровня шума замены

F i g. 1. Dependency graph of the proportion of correctly recognized periods to the replacement noise



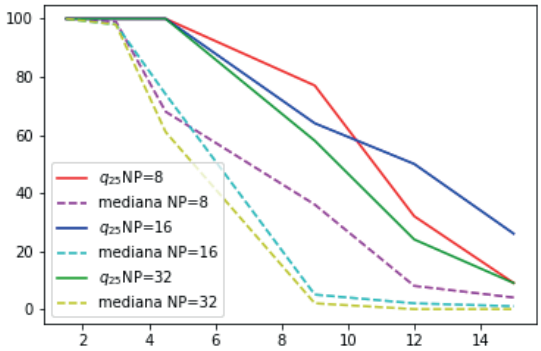
Р и с. 2. График зависимости доли правильно распознанных периодов от уровня шума удаления

F i g. 2. Dependency graph of the proportion of correctly recognized periods to the removal noise



Р и с. 3. График зависимости доли правильно распознанных периодов от уровня шума вставки

F i g. 3. Dependency graph of the proportion of correctly recognized periods to the insert noise



Р и с. 4. График зависимости доли правильно распознанных периодов от уровня смешанного шума

F i g. 4. Dependency graph of the proportion of correctly recognized periods to the mixed noise level

На графиках пунктирной линией изображена доля искаженных последовательностей, в которых медиана с точностью до 10% оказалась равна периоду p. Сплошная линия обозначает долю искаженных последовательностей, в которой с периодом исходной последовательности (с точностью до 10%) совпадает квартиль q_{25}. Из рисунков видно, что квартиль является более устойчивой к шуму оценкой периода, чем медиана. Кроме того, метод позволяет удовлетворительно определять период при уровне шума до 5%. При уровне шума от 5% до 10% период определяется нестабильно, т.е. для одних искаженных вариантов одной и той же периодической последовательности получаются достаточно точные оценки периода, а для других метод дает завышение обычно в два раза, очень редко в три раза.

Отметим также, что качество идентификации длины периода зависит от числа периодов, и увеличение числа периодов приводит к лучшим результатам при использовании квантили. Анализ экспериментальных результатов также показывает, что предложенный метод дает сопоставимые результаты как для шумов различных типов, так и для смешанного шума.

Список использованных источников

[1] Zhukova, G. Informative Symbolic Representations as a Way to Qualitatively Analyze Time Series / G. Zhukova, Y. Smetanin, M. Uljanov. – DOI 10.1109/EnT.2019.00014 // 2019 International Conference on Engineering Technologies and Computer Science (EnT). – Moscow, Russia, 2019. – Pp. 43-47. – URL: https://ieeexplore.ieee.org/document/8711889 (дата обращения: 15.01.2020).

[2] Lin, J. A symbolic representation of time series, with implications for streaming algorithms / J. Lin, E. Keogh, S. Lonardi, B. Chiu. – DOI 10.1145/882082.882086 // Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery (DMKD '03). – Association for Computing Machinery, New York, NY, USA, 2003. – Pp. 2-11. – URL: http://www.cs.ucr.edu/~eamonn/SAX.pdf (дата обращения: 15.01.2020).

[3] He, Z. Mining partial periodic correlations in time series / Z. He, X. S. Wang, B. S. Lee, A. C. H. Ling. – DOI 10.1007/s10115-006-0051-5 // Knowledge and Information Systems. – 2008. – Vol. 15, issue 1. – Pp. 31-54. – URL: https://link.springer.com/article/10.1007/s10115-006-0051-5 (дата обращения: 15.01.2020).

[4] Chanda, A. K. A new framework for mining weighted periodic patterns in time series databases / A. K. Chanda, C. F. Ahmed, Md. Samiullah, C. K. Leung. – DOI 10.1016/j.eswa.2017.02.028 // Expert Systems with Applications. – 2017. – Vol. 79, no. C. – Pp. 207-224. – URL: https://www.sciencedirect.com/science/article/pii/S0957417417301173 (дата обращения: 15.01.2020).

[5] Yang, K.-J. A two-phase approach for mining weighted partial periodic patterns / K.-J.Yang, T.-P. Hong, G.-C. Lan, Y.-M. Chen. – DOI 10.1016/j.engappai.2014.01.004 // Engineering Applications of Artificial Intelligence. – 2014. – Vol. 30. – Pp. 225-234. – URL: https://www.sciencedirect.com/science/article/pii/S0952197614000116 (дата обращения: 15.01.2020).

[6] He, R. Automated Mining of Approximate Periodicity on Numeric Data: A Statistical Approach / R. He, S. Yang, J. Yang, J. Cao. – DOI 10.1145/3193077.3194509 // Proceedings of the 2nd International Conference on Compute and Data Analysis (ICCDA 2018). – ACM, New York, NY, USA, 2018. – Pp. 20-27. – URL: https://dl.acm.org/doi/abs/10.1145/3193077.3194509 (дата обращения: 15.01.2020).

[7] Patel, M. A Comprehensive Study on Periodicity Mining Algorithms / M. Patel, N. Modi. – DOI 10.1109/ICGTSPICC.2016.7955365 // 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC). – Jalgaon, 2016. – Pp. 567-575. – URL: https://ieeexplore.ieee.org/document/7955365 (дата обращения: 15.01.2020).

[8] Aydin, B. Spatiotemporal Frequent Pattern Mining on Solar Data: Current Algorithms and Future Directions / B. Aydin, R. Angryk. – DOI 10.1109/ICDMW.2015.10 // 2015 IEEE International Conference on Data Mining Workshop (ICDMW). – Atlantic City, NJ, 2015. – Pp. 575-581. – URL: https://ieeexplore.ieee.org/document/7395719 (дата обращения: 15.01.2020).

[9] Dong, S. An Innovative Model to Mine Asynchronous Periodic Pattern of Moving Objects / S. Dong, S. Liu, Y. Zhao, Z. Shao. – DOI 10.1007/s11042-018-6752-4 // Multimedia Tools and Applications. – 2019. – Vol. 78, issue 7. – Pp. 8943-8964. – URL: https://link.springer.com/article/10.1007/s11042-018-6752-4 (дата обращения: 15.01.2020).

[10] Bjørnstad, O. N. Timing and Periodicity of Influenza Epidemics / O. N. Bjørnstad, C. Viboud. – DOI 10.1073/pnas.1616052113 // Proceedings of the National Academy of Sciences. – 2016. – Vol. 113, no. 46. – Pp. 12899-12901. – URL: https://www.pnas.org/content/113/46/12899 (дата обращения: 15.01.2020).

[11] Parthasarathy, S. Robust periodicity detection algorithms / S. Parthasarathy, S. Mehta, S. Srinivasan. – DOI 10.1145/1183614.1183774 // Proceedings of the 15th ACM international conference on Information and knowledge management (CIKM '06). – Association for Computing Machinery, New York, NY, USA, 2006. – Pp. 874-875. – URL: https://dl.acm.org/doi/10.1145/1183614.1183774 (дата обращения: 15.01.2020).

[12] Otunba, R. APT: Approximate Period Detection in Time Series / R. Otunba, J. Lin // Proceedings of the 26th International Conference on Software Engineering & Knowledge Engineering (SEKE). – Vancouver, Canada, 2014. – Pp. 490-494. – URL: https://ksiresearchorg.ipage.com/seke/seke14paper/seke14paper_9.pdf (дата обращения: 15.01.2020).

[13] Vlachos, M. On Periodicity Detection and Structural Periodic Similarity / M. Vlachos, P. Yu, V. Castelli. – DOI 10.1137/1.9781611972757.40 // Proceedings of the 2005 SIAM International Conference on Data Mining. – Society for Industrial and Applied Mathematics, 2005. – Pp. 449-460. – URL: https://epubs.siam.org/doi/10.1137/1.9781611972757.40 (дата обращения: 15.01.2020).

[14] Korotkov, E. V. Latent periodicity of DNA sequences from some human gene regions / E. V. Korotkov, M. A. Korotkova. – DOI 10.3109/10425179509020866 // DNA Sequence. – 1995. – Vol. 5, issue 6. – Pp. 353-358. – URL: https://www.tandfonline.com/doi/abs/10.3109/10425179509020866 (дата обращения: 15.01.2020).

[15] Нестеренко, А. Ю. Алгоритмы поиска длин циклов в последовательностях и их приложения / А. Ю. Нестеренко // Фундаментальная и прикладная математика. – 2010. – Т. 16, № 6. – С. 109-122. – URL: https://elibrary.ru/item.asp?id=20285258 (дата обращения: 15.01.2020).

[16] Elfeky, M. G. WARP: time warping for periodicity detection / M. G. Elfeky, W. G. Aref, A. K. Elmagarmid. – DOI 10.1109/ICDM.2005.152 // Fifth IEEE International Conference on Data Mining (ICDM'05). – Houston, TX, 2005. – Pp. 8. – URL: https://ieeexplore.ieee.org/document/1565672 (дата обращения: 15.01.2020).

[17] Elfeky, M.G. Periodicity detection in time series databases / M. G. Elfeky, W. G. Aref, A. K. Elmagarmid. – DOI 10.1109/TKDE.2005.114 // IEEE Transactions on Knowledge and Data Engineering. – 2005. – Vol. 17, issue 7. – Pp. 875-887. – URL: https://ieeexplore.ieee.org/document/1432698 (дата обращения: 15.01.2020).

[18] Rasheed, F. STNR: A suffix tree based noise resilient algorithm for periodicity detection in time series databases / F. Rasheed, R. Alhajj. – DOI 10.1007/s10489-008-0144-9 // Applied Intelligence. – 2010. – Vol. 32, issue 3. – Pp. 267-278. – URL: https://link.springer.com/article/10.1007%2Fs10489-008-0144-9 (дата обращения: 15.01.2020).

[19] Ukkonen, E. On-line construction of suffix trees / E. Ukkonen. – DOI 10.1007/BF01206331 // Algorithmica. – 1995. – Vol. 14, issue 3. – Pp. 249-260. – URL: https://link.springer.com/article/10.1007%2FBF01206331 (дата обращения: 15.01.2020).

[20] Korotkov, E. V. Developing New Mathematical Method for Search of the Time Series Periodicity with Deletions and Insertions / E. V. Korotkov, M. A. Korotkova. – DOI 10.1088/1742-6596/788/1/012019 // Journal of Physics: Conference Series. – 2017. – Vol. 788, issue 1. – Article number: 012019. – URL: https://iopscience.iop.org/article/10.1088/1742-6596/788/1/012019 (дата обращения: 15.01.2020).

[21] Frenkel, F. E. Database of Periodic DNA Regions in Major Genomes / F. E. Frenkel, M. A. Korotkova, E. V. Korotkov. – DOI 10.1155/2017/7949287 // BioMed Research International. – 2017. – Vol. 2017. – Article number: 7949287. – URL: https://www.hindawi.com/journals/bmri/2017/7949287/ (дата обращения: 15.01.2020).

[22] Chaley, M. B. Method Revealing Latent Periodicity of the Nucleotide Sequences Modified for a Case of Small Samples / M. B. Chaley, E. V. Korotkov, K. G. Skryabin. – DOI 10.1093/dnares/6.3.153 // DNA Research. – 1999. – Vol. 6, issue 3. – Pp. 153-163. – URL: https://academic.oup.com/dnaresearch/article/6/3/153/421488 (дата обращения: 15.01.2020).

[23] Жукова, Г. Н. Сравнение основных алгоритмов поиска циклов в символьных последовательностях при наличии искажений / Г. Н. Жукова, Ю. Г. Сметанин, М. В. Ульянов. – DOI 10.25559/SITITO.15.201904.905-915 // Современные информационные технологии и ИТ-образование. – 2019. – Т. 15, № 4. – С. 905-915. – URL: https://elibrary.ru/item.asp?id=43575610 (дата обращения: 15.01.2020).

[24] Сметанин, Ю. Г. Энтропийный подход к построению меры символьного разнообразия слов и его применение к кластеризации геномов растений / Ю. Г. Сметанин, М. В. Ульянов, А. С. Пестова. – DOI 10.17537/2016.11.114 // Математическая биология и биоинформатика. – 2016. – Т. 11, № 1. – С. 114-126. – URL: https://elibrary.ru/item.asp?id=28120766 (дата обращения: 15.01.2020).

[25] Жукова, Г. Н. Вероятностная модель шумов для периодических символьных последовательностей / Г. Н. Жукова, Ю. Г. Сметанин, М. В. Ульянов. – DOI 10.25559/SITITO.15.201902.431-440 // Современные информационные технологии и ИТ-образование. – 2019. – Т. 15, № 2. – С. 431-440. – URL: https://elibrary.ru/item.asp?id=41244513 (дата обращения: 15.01.2020).

Поступила 15.01.2020; принята к публикации 18.03.2020; опубликована онлайн 25.05.2020.

Об авторах:

Жукова Галина Николаевна, доцент Департамента программной инженерии, Факультет компьютерных наук, Национальный исследовательский университет «Высшая школа экономики» (101000, Россия, г. Москва, ул. Мясницкая, д. 20), кандидат физико-математических наук, доцент, ORCID: http://orcid.org/0000-0003-1835-7422, gzhukova@hse.ru

Жуков Алексей Владимирович, ведущий инженер-программист, Национальный исследовательский центр «Курчатовский институт» (123182, Россия, г. Москва, пл. Академика Курчатова, д. 1), ORCID: http://orcid.org/0000-0002-7266-467X, zhukov_avl@nrcki.ru

Сметанин Юрий Геннадиевич, главный научный сотрудник, Вычислительный центр им. А.А. Дородницына РАН, Федеральный исследовательский центр «Информатика и управление» Российской академии наук (119333, Россия, г. Москва, ул. Вавилова, д. 44, корп. 2), доктор физико-математических наук, ORCID: http://orcid.org/0000-0003-0242-6972, ysmetanin@rambler.ru

Ульянов Михаил Васильевич, ведущий научный сотрудник, Институт проблем управления им. В.А. Трапезникова Российской академии наук (117997, Россия, г. Москва, ул. Профсоюзная, д. 65), профессор кафедры алгоритмических языков, факультет вычислительной математики и кибернетики, Московский государственный университет имени М.В. Ломоносова (119991, Россия, г. Москва, ГСП-1, Ленинские горы, д. 1), доктор технических наук, профессор, ORCID: http://orcid.org/0000-0002-5784-9836, muljanov@mail.ru

Все авторы прочитали и одобрили окончательный вариант рукописи.

References

[1] Zhukova G., Smetanin Y., Uljanov M. Informative Symbolic Representations as a Way to Qualitatively Analyze Time Series. In: 2019 International Conference on Engineering Technologies and Computer Science (EnT). Moscow, Russia; 2019. p. 43-47. (In Eng.) DOI: https://doi.org/10.1109/EnT.2019.00014

[2] Lin J., Keogh E., Lonardi S., Chiu B. A symbolic representation of time series, with implications for streaming algorithms. In: Proceedings of the 8th ACM SIGMOD workshop

on Research issues in data mining and knowledge discovery (DMKD '03). Association for Computing Machinery, New York, NY, USA; 2003. p. 2-11. (In Eng.) DOI: https://doi.org/10.1145/882082.882086

[3] He Z., Wang X.S., Lee B.S., Ling A.C.H. Mining partial periodic correlations in time series. *Knowledge and Information Systems.* 2008; 15(1):31-54. (In Eng.) DOI: https://doi.org/10.1007/s10115-006-0051-5

[4] Chanda A.K., Ahmed C.F., Samiullah Md., Leung C.K.A new framework for mining weighted periodic patterns in time series databases. *Expert Systems with Applications*. 2017; 79(C):207-224. (In Eng.) DOI: https://doi.org/10.1016/j.eswa.2017.02.028

[5] Yang K.-J., Hong T.-P., Lan G.-C., Chen Y.-M. A two-phase approach for mining weighted partial periodic patterns. *Engineering Applications of Artificial Intelligence*. 2014; 30:225-234. (In Eng.) DOI: https://doi.org/10.1016/j.engappai.2014.01.004

[6] He R., Yang S., Yang J., Cao J. Automated Mining of Approximate Periodicity on Numeric Data: A Statistical Approach. In: Proceedings of the 2nd International Conference on Compute and Data Analysis (ICCDA 2018). ACM, New York, NY, USA; 2018. p. 20-27. (In Eng.) DOI: https://doi.org/10.1145/3193077.3194509

[7] Patel M., Modi N. A Comprehensive Study on Periodicity Mining Algorithms. In: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC). Jalgaon; 2016. p. 567-575. (In Eng.) DOI: https://doi.org/10.1109/ICGTSPICC.2016.7955365

[8] Aydin B., Angryk R. Spatiotemporal Frequent Pattern Mining on Solar Data: Current Algorithms and Future Directions. In: 2015 IEEE International Conference on Data Mining Workshop (ICDMW). Atlantic City, NJ; 2015. p. 575-581. (In Eng.) DOI: https://doi.org/10.1109/ICDMW.2015.10

[9] Dong S., Liu S., Zhao Y., Shao Z. An Innovative Model to Mine Asynchronous Periodic Pattern of Moving Objects. *Multimedia Tools and Applications*. 2019; 78(7):8943-8964. (In Eng.) DOI: https://doi.org/10.1007/s11042-018-6752-4

[10] Bjørnstad O.N., Viboud C. Timing and Periodicity of Influenza Epidemics. *Proceedings of the National Academy of Sciences.* 2016; 113(46):12899-12901. (In Eng.) DOI: https://doi.org/10.1073/pnas.1616052113

[11] Parthasarathy S., Mehta S., Srinivasan S. Robust periodicity detection algorithms. In: Proceedings of the 15th ACM international conference on Information and knowledge management (CIKM '06). Association for Computing Machinery, New York, NY, USA; 2006. p. 874-875. (In Eng.) DOI: https://doi.org/10.1145/1183614.1183774

[12] Otunba R., Lin J. APT: Approximate Period Detection in Time Series. In: Proceedings of the 26th International Conference on Software Engineering & Knowledge Engineering (SEKE). Vancouver, Canada; 2014. p. 490-494. Available at: https://ksiresearchorg.ipage.com/seke/seke14paper/seke14paper_9.pdf (accessed 15.01.2020). (In Eng.)

[13] Vlachos M., Yu P., Castelli V. On Periodicity Detection and Structural Periodic Similarity. In: Proceedings of the 2005 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics; 2005. p. 449-460. (In Eng.) DOI: https://doi.org/10.1137/1.9781611972757.40

[14] Korotkov E.V., Korotkova M.A. Latent periodicity of DNA sequences from some human gene regions. *DNA Sequence.* 1995; 5(6):353-358. (In Eng.) DOI: https://doi.org/10.3109/10425179509020866

[15] Nesterenko A.Yu. Cycle detection algorithms and their applications. *Fundamentalnaya i prikladnaya matematika* = Fundamental and Applied Mathematics. 2010; 16(6):109-122. Available at: https://elibrary.ru/item.asp?id=20285258 (accessed 15.01.2020). (In Russ., abstract in Eng.)

[16] Elfeky M.G., Aref W.G., Elmagarmid A.K. WARP: time warping for periodicity detection. In: Fifth IEEE International Conference on Data Mining (ICDM'05). Houston, TX; 2005. p. 8. (In Eng.) DOI: https://doi.org/10.1109/ICDM.2005.152

[17] Elfeky M.G., Aref W.G., Elmagarmid A.K. Periodicity detection in time series databases. *IEEE Transactions on Knowledge and Data Engineering.* 2005; 17(7):875-887. (In Eng.) DOI: https://doi.org/10.1109/TKDE.2005.114

[18] Rasheed F., Alhajj R. STNR: A suffix tree based noise resilient algorithm for periodicity detection in time series databases. *Applied Intelligence.* 2010; 32(3):267-278. (In Eng.) DOI: https://doi.org/10.1007/s10489-008-0144-9

[19] Ukkonen E. On-line construction of suffix trees. *Algorithmica.* 1995; 14(3):249-260. (In Eng.) DOI: https://doi.org/10.1007/BF01206331

[20] Korotkov E.V., Korotkova M.A. Developing New Mathematical Method for Search of the Time Series Periodicity with Deletions and Insertions. *Journal of Physics: Conference Series.* 2017; 788(1):012019. (In Eng.) DOI: https://doi.org/10.1088/1742-6596/788/1/012019

[21] Frenkel F.E., Korotkova M.A., Korotkov E.V. Database of Periodic DNA Regions in Major Genomes. *BioMed Research International.* 2017; 2017:7949287. (In Eng.) DOI: https://doi.org/10.1155/2017/7949287

[22] Chaley M.B., Korotkov E.V., Skryabin K.G. Method Revealing Latent Periodicity of the Nucleotide Sequences Modified for a Case of Small Samples. *DNA Research.* 1999; 6(3):153-163. (In Eng.) DOI: https://doi.org/10.1093/dnares/6.3.153

[23] Zhukova G.N., Smetanin Yu.G., Ulyanov M.V. Comparison of Some Algorithms for Periodicity Detection in Symbolic Sequences in the Presence of Distortions. *Sovremennye informacionnye tehnologii i IT-obrazovanie* = Modern Information Technologies and IT-Education. 2019; 15(4):905-915. (In Russ., abstract in Eng.) DOI: https://doi.org/10.25559/SITITO.15.201904.905-915

[24] Smetanin Y.G., Ulyanov M.V., Pestova A.S. Entropy Approach to the Construction of a Measure of Word Symbolic Diverseness and its Application to Clustering of Plant Genomes. *Mathematical Biology and Bioinformatics.* 2016; 11(1):114-126. (In Russ., abstract in Eng.) DOI: https://doi.org/10.17537/2016.11.114

[25] Zhukova G.N., Smetanin Yu.G., Ulyanov M.V. A Stochastic Model of Noises for Periodic Symbol Sequences. *Sovremennye informacionnye tehnologii i IT-obrazovanie* = Modern Information Technologies and IT-Education. 2019; 15(2):431-440. (In Russ., abstract in Eng.) DOI: https://doi.org/10.25559/SITITO.15.201902.431-440

Submitted 15.01.2020; revised 18.03.2020; published online 25.05.2020.

About the authors:

Galina N. Zhukova, Associate Professor of the School of Software Engineering, Faculty of Computer Science, National Research University Higher School of Economics (20 Myasnitskaya St., Moscow 101000, Russia), Ph.D. (Phys.-Math.), Associate Professor, ORCID: http://orcid.org/0000-0003-1835-7422, gzhukova@hse.ru

Alexey V. Zhukov, Senior Software Developer, National Research Center "Kurchatov Institute" (1 Akademika Kurchatova pl., Moscow 123182, Russia), ORCID: http://orcid.org/0000-0002-7266-467X, zhukov_avl@nrcki.ru

Yuri G. Smetanin, Chief Researcher of the Dorodnicyn Computing Centre of RAS, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (44-2 Vavilov St., Moscow 119333, Russia), Dr. Sci. (Phys.-Math.), ORCID: http://orcid.org/0000-0003-0242-6972, ysmetanin@rambler.ru

Mikhail V. Uljanov, Leading Scientist Researcher, V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences (65 Profsoyuznaya St., Moscow 117997, Russia), Professor of the Department of Algorithmic Languages, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University (1, Leninskie gory, Moscow 119991, Russia), Dr.Sci. (Engineering), Professor, ORCID: http://orcid.org/0000-0002-5784-9836, muljanov@mail.ru

All authors have read and approved the final manuscript.