Применение параллельных вычислений для аннотирования сенсорных данных*

М.Л. Цымблер, А.И. Гоглачев

Южно-Уральский государственный университет (Челябинск)

Аннотирование сенсорных данных предполагает автоматизированную разметку временного ряда показаний, снятых с сенсора, которая выделяет различные активности, заданные указанным рядом. Разметка активностей имеет широкий спектр практического применения: предиктивное техническое обслуживание, умное управление системами жизнеобеспечения, моделирование климата и др. Ранее нами разработан параллельный алгоритм PSF для аннотирования данных сенсоров с помощью графического процессора. В данной статье описаны два тематических исследования, выполненные с помощью алгоритма PSF: аннотирование показаний носимого виброакселерометра, закрепленного на человеке, и стационарного виброакселерометра, установленного на малогабаритной дробильной установке.

Ключевые слова: временной ряд, аннотирование, сниппет, параллельный алгоритм, графический процессор.

1. Введение

Аннотирование сенсорных данных предполагает автоматизированную разметку временного ряда показаний, снятых с сенсора, которая выделяет различные активности, заданные указанным рядом. Разметка активностей позволяет кратко описать и визуализировать сенсорные данные и поэтому имеет широкий спектр практического применения: предиктивное техническое обслуживание в цифровой индустрии, умное управление системами жизнеобеспечения, мониторинг показателей функциональной диагностики организма человека, моделирование климата и др.

Для решения задачи аннотирования предложены различные подходы: лейтмотивы [2], шейпелеты [3], ослабленные периоды и средние тенденции [4] и др. Указанные подходы, однако, не являются независимыми от предметной области либо не обеспечивают количественную оценку покрытия выделенных активностей. Например, лейтмотив представляет собой пару наиболее похожих друг на друга подпоследовательностей временного ряда, но доля ряда, покрываемая таким шаблоном, неизвестна. Шейплет определяется как подпоследовательность, одновременно наиболее похожая на большинство подпоследовательностей данного класса и наиболее отличающаяся от подпоследовательностей из других классов. Шейплеты допускают количественную оценку покрытия, однако требуют знаний о предметной области.

В недавней работе [1] предложена концепция сниппета (snippet), которая свободна от указанных выше недостатков. Сниппет представляет собой подпоследовательность заданной длины, на которую похожи многие другие подпоследовательности данного ряда в смысле специальной меры схожести [5]. Эксперименты показывают, что сниппеты позволяют адекватно аннотировать сенсорные данные из широкого спектра предметных областей [1]. Однако оригинальный алгоритм поиска сниппетов имеет высокую вычислительную сложность, что критично при обработке временных рядов, насчитывающих от сотни тысяч элементов.

В нашей предыдущей работе [8] предложен параллельный алгоритм поиска сниппе-

*Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (грант № 20-07-00140) и Министерства науки и высшего образования РФ (государственное задание FENU-2020-0022).

тов для графического процессора, названный PSF. В данной статье продолжается начатая работа и описаны два тематических исследования по аннотированию сенсорных данных, выполненные с помощью алгоритма PSF: анализ показаний носимого виброакселерометра, закрепленного на человеке, и стационарного виброакселерометра, установленного на малогабаритной дробильной установке. Остаток статьи имеет следующую структуру. Раздел 2 содержит формальные определения и описание алгоритма PSF. В разделе 3 описаны проведенные исследования. Заключение резюмирует полученные результаты.

2. Параллельный алгоритм поиска активностей

2.1. Формальные определения и обозначения

Временной ряд (time series) T представляет собой последовательность хронологически упорядоченных вещественных значений:

$$T = (t_1, \ldots, t_n),\ t_i \in \mathbb{R} \tag{1}$$

Число n обозначается как $|T|$ и называется длиной ряда.

Подпоследовательность (subsequence) $T_{i,m}$ временного ряда T представляет собой непрерывное подмножество T из m элементов, начиная с позиции i:

$$T_{i,m} = (t_i, \ldots, t_{i+m-1}),\ 1 \leqslant m \ll n, 1 \leqslant i \leqslant n - m + 1. \tag{2}$$

Временной ряд T может быть логически разбит на сегменты – непересекающиеся подпоследовательности заданной длины m. Здесь и далее без существенного ограничения общности мы можем считать, что n кратно m, поскольку $m \ll n$ Множество сегментов ряда, имеющих длину $m \ll n$, обозначим как S_T^m, элементы этого множества как $S_1, \ldots, S_{n/m}$:

$$S_T^m = (S_1, \ldots, S_{n/m}),\ S_i = T_{m\cdot(i-1)+1,\,m}. \tag{3}$$

Концепция *сниппетов (snippet)* предложена Кеогом и др. в работе [1] и уточняет понятие типичных подпоследовательностей временного ряда следующим образом. Каждый сниппет представляет собой один из сегментов временного ряда. Со сниппетом ассоциируются его ближайшие соседи – подпоследовательности ряда, имеющие ту же длину, что и сниппет, которые более похожи на данный сниппет, чем на другие сегменты. Для вычисления схожести подпоследовательностей используется специализированная мера схожести MPdist, основанная на евклидовом расстоянии. Сниппеты упорядочиваются по убыванию мощности множества своих ближайших соседей. Множество сниппетов ряда T, имеющих длину m обозначается, как C_T^m, а элементы этого множества – как $C_1, \ldots, C_K$:

$$C_T^m = (C_1, \ldots, C_K),\ C_i \in S_T^m. \tag{4}$$

Число K $(1 \leqslant K \leqslant n/m)$ представляет собой параметр, задаваемый прикладным программистом, и отражает соответствующее количество наиболее типичных сниппетов. С каждым сниппетом ассоциированы следующие атрибуты: индекс сниппета, ближайшие соседи и значимость данного сниппета. Сниппеты упорядочиваются по убыванию их значимости.

Мера MPdist [5], используемая для вычисления схожести подпоследовательностей при нахождении сниппетов, неформально определяется следующим образом. Два временных ряда равной длины m тем более похожи друг на друга в смысле меры MPdist, чем больше в каждом из них имеется подпоследовательностей заданной длины ℓ $(3 \leqslant \ell \leqslant m)$, близких друг к другу в смысле нормализованного евклидова расстояния.

2.2. Реализация

В данном разделе приводится краткое описание алгоритма PSF, предложенного в нашей предыдущей работе [8].

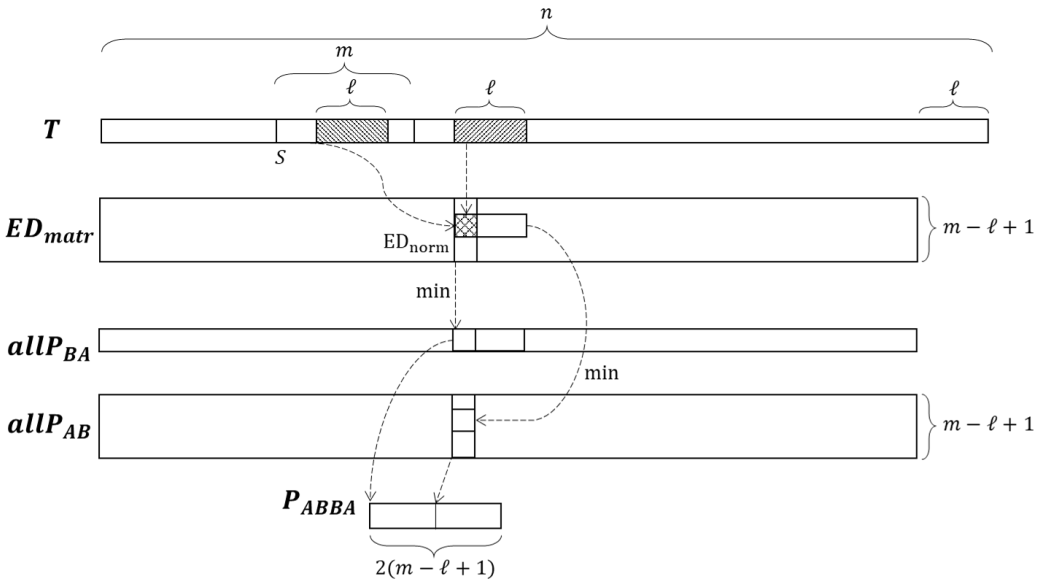


Рис. 1. Структуры данных алгоритма PSF

Структуры данных алгоритма PSF представлены на рис. 1. Ключевой для распараллеливания структурой данных является матрица ED_{norm}-расстояний между каждой подпоследовательностью длины ℓ сегмента S и каждой подпоследовательностью длины ℓ исходного ряда. Обозначим указанную матрицу за ED_{matr}:

$$ED_{matr} \in \mathbb{R}^{(m-\ell+1)\times(n-\ell+1)} : ED_{matr}(i,j) = \mathrm{ED}_{\mathrm{norm}}(S_{i,\ell}, T_{j,\ell}). \tag{5}$$

Параллелизм вычислений матрицы расстояний ED_{matr} реализован на основе следующей техники, предложенной в работе [7]. Сначала вычисляется матрица центрированных сумм произведений значений ряда, которая используется для вычисления корреляции по Пирсону между подпоследовательностями ряда. Далее значения корреляции по Пирсону между двумя подпоследовательностями ряда преобразуюся в z-нормализованное евклидово расстояние.

На втором шаге в каждом столбце матрицы ED_{matr}, полученной на первом шаге, находится минимум. Обозначим вектор таких минимумов за $allP_{BA}$:

$$allP_{BA} \in \mathbb{R}^{n-\ell+1} : allP_{BA}(j) = \min_{1\leqslant i\leqslant m-\ell+1} ED_{matr}(i,j). \tag{6}$$

На третьем шаге в каждой строке ED_{matr} выполняется поиск минимумов в скользящем окне длины ℓ. Обозначим матрицу таких минимумов за $allP_{AB}$:

$$allP_{AB} \in \mathbb{R}^{(m-\ell+1)\times(n-\ell+1)} : allP_{AB}(i,j) = \min_{j\leqslant c\leqslant j+m-\ell+1} ED_{matr}(i,c). \tag{7}$$

На четвертом шаге для каждой подпоследовательности ряда, имеющей длину ℓ, и сегмента S выполняется построение матричного профиля. Для построения одного матричного профиля выполняется сцепление соответствующих данной подпоследовательности столбца матрицы $allP_{AB}$ и подпоследовательности длины $m-\ell+1$, входящей в вектор $allP_{BA}$. Результат сцепления обозначим как вектор P_{ABBA}:

$$P_{ABBA} \in \mathbb{R}^{2(m-\ell+1)} : P_{ABBA}(T_{j,\ell}) = allP_{AB}(1,j) \odot \ldots \odot allP_{AB}(m-\ell+1,j) \odot \\ \odot allP_{BA}(j) \odot \ldots \odot allP_{BA}(m-\ell+1), \tag{8}$$

где $1 \leqslant j \leqslant n-\ell+1$. Для финального вычисления меры схожести MPdist между сегментом и подпоследовательностью необходимо выполнить сортировку P_{ABBA} и взять k-е значение упорядоченного массива).

3. Вычислительные эксперименты

В данном разделе описаны тематические исследования по применению параллельного алгоритма PSF для аннотирования сенсорных данных. Вычислительные эксперименты были проведены на следующей аппаратной платформе: центральный процессор Intel Xeon Gold 6254, 18 ядер, тактовая частота 4.0 GHz, графический процессор NVIDIA Tesla V100 SXM2, количество ядер 5120, тактовая частота 1.3 GHz, пиковая производительность 15.7 TFLOPS.

Для оценки эффективности аннотирования нами используются стандартные меры качества классификации, определяемые следующим образом:

$$\text{Precision} = \frac{TP}{TP+FP}, \ \text{Recall} = \frac{TP}{TP+FN}, \ \text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{9}$$

где TP, FP, TN и FN – количество истинно-положительных, ложно-положительных, истинно-отрицательных и ложно-отрицательных элементов ряда соответственно.

3.1. Аннотирование сенсорных данных носимого акселерометра

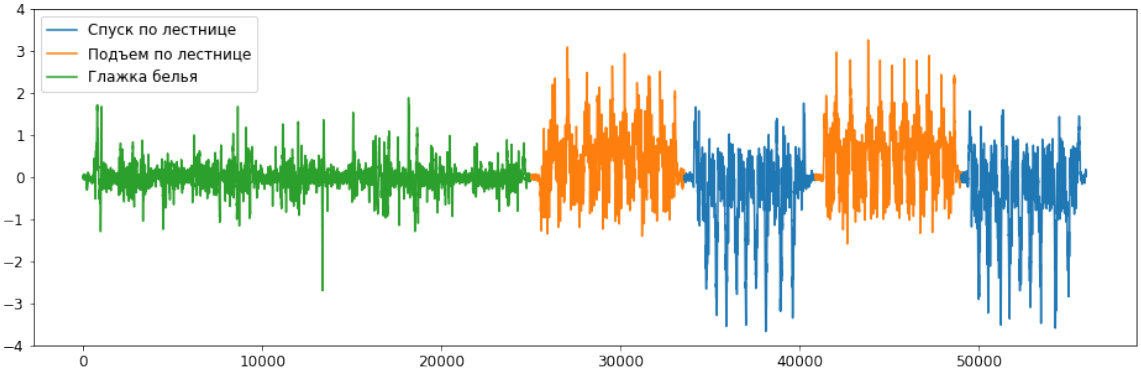
Для первого тематического исследования нами взят отрезок временного ряда PAMAP [6], представляющий собой показания закрепленного на человеке виброакселерометра, для которых известны виды физической активности, выполнявшиеся этим человеком. Данный ряд содержит три вида физической активности: глажка белья, подъем по лестнице, спуск по лестнице. Количество сниппетов соответствовало числу активностей, отражаемых временным рядом, т.е. $K = 3$. Длина сегмента $m = 800$. В табл. 1 представлена оценка эффективности аннотирования по мерам, указанным в формуле (9).

Таблица 1. Показатели качества аннотирования

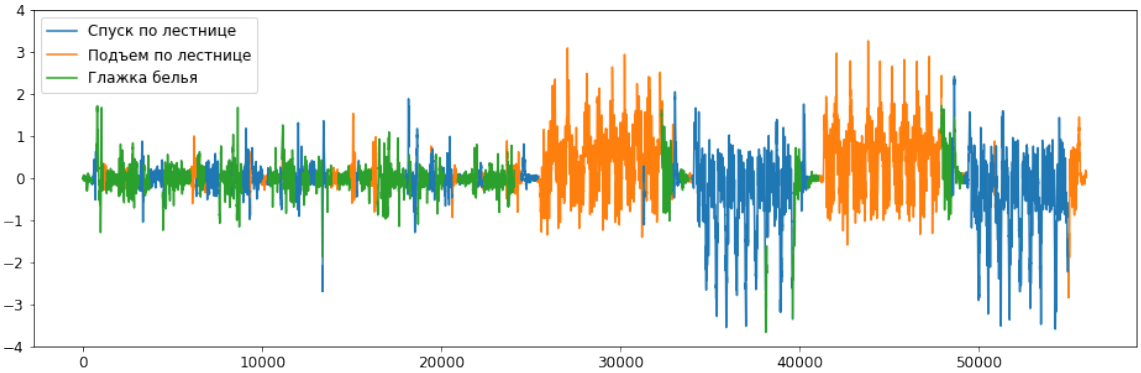
Активность	Точность	Полнота	F1-мера
Подъем по лестнице	0.74	0.83	0.78
Спуск по лестнице	0.59	0.82	0.68
Глажка белья	0.83	0.58	0.68

Из табл. 1 видно, что наилучшим образом распознается подъем по лестнице, наихудшим – глажка белья. Можно видеть высокую полноту при распозанвании подъема и спуска по лестнице и меньшую – при глажке белья. На рис. 2 представлены исходная разметка ряда и результат аннотирования ряда при помощи алгоритма PSF.

На рис. 3 представлены найденные сниппеты, соответствующие активностям на временном ряду. Числами на левой стороне рисунка указаны индексы начала сниппетов.



а) исходная разметка ряда



б) результат работы PSF

Рис. 2. Временной ряд PAMAP

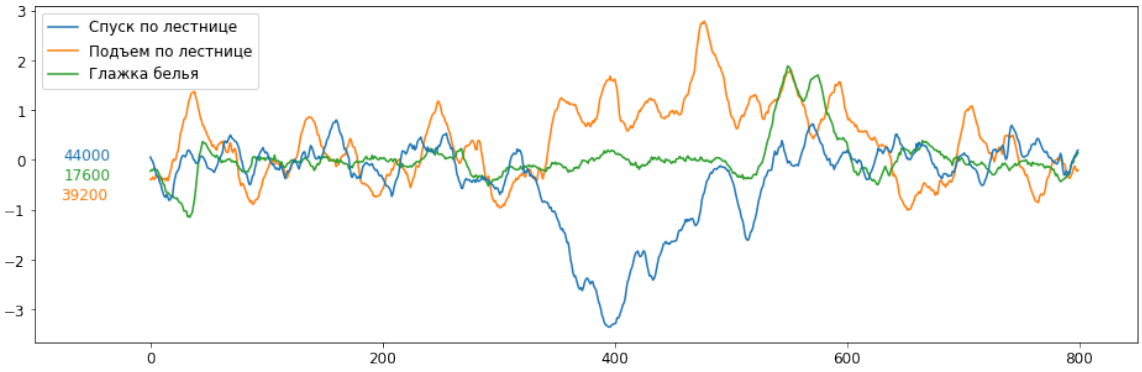


Рис. 3. Найденные сниппеты

3.2. Аннотирование показаний сенсора, установленного на промышленном оборудовании

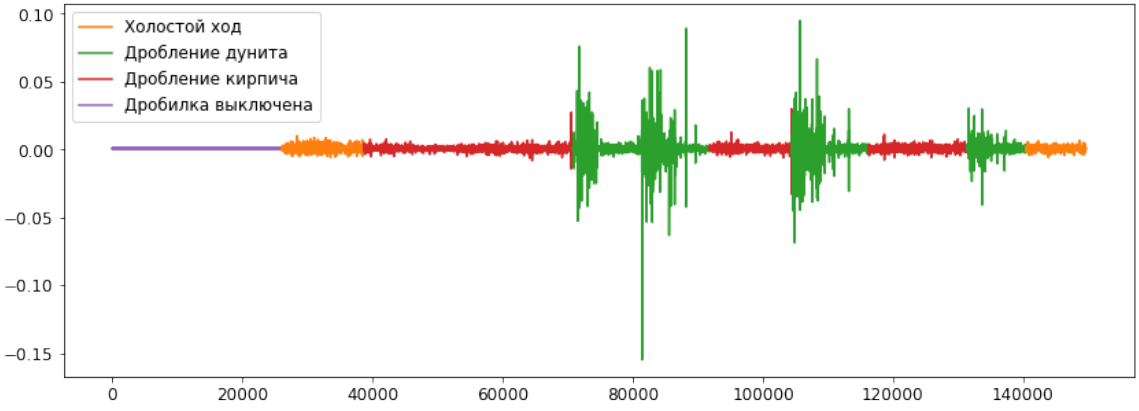
Для второго тематического исследования нами используются данные виброакселерометра, установленного на малогабаритной дробильной установке. Показания записаны во время заброса двух материалов различной твердости: дунита и кирпича. Помимо дробления указанных материалов, записаны два других вида активности: установка выключена и холостой ход. Количество сниппетов соответствует общему числу активностей: $K = 4$. Длина сегмента соответствует пяти секундам: $m = 4000$. В табл. 2 представлена оценка эффективности аннотирования по мерам, указанным в формуле (9).

По данным в табл. 2 можно видеть, что дробление дунита и выключенное состояние

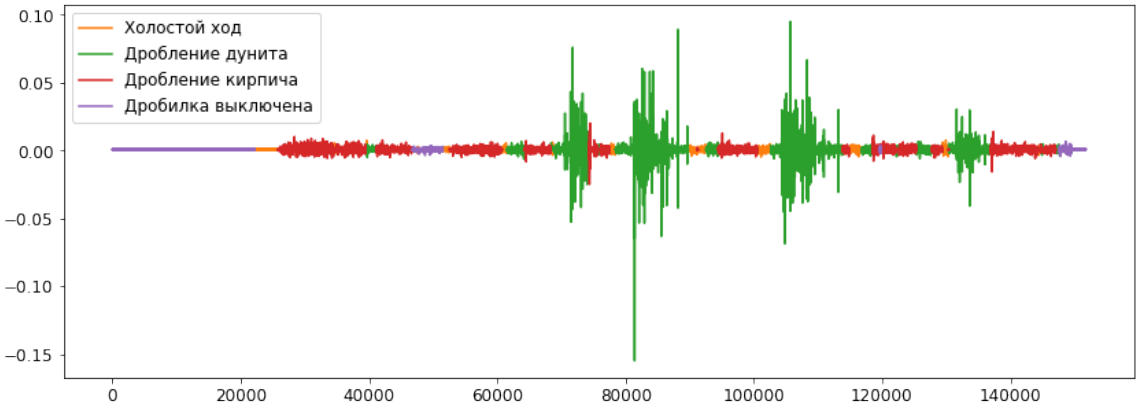
Таблица 2. Показатели качества аннотирования

Активность	Точность	Полнота	F1-мера
Холостой ход	0.05	0.03	0.04
Дробление дунита	0.65	0.72	0.68
Дробление кирпича	0.55	0.54	0.55
Установка выключена	0.77	0.85	0.81

дробилки имеют наиболее высокие точность и полноту распознавания. Интегрально наилучшим образом распознается дробление дунита, наихудшим – холостой ход. На рис. 4 представлены исходная разметка ряда и результат аннотирования ряда при помощи алгоритма PSF.



а) исходная разметка ряда



б) результат работы PSF

Рис. 4. Временной ряд показаний сенсора, установленного на дробильной установке

На рис. 5 представлены найденные сниппеты.

4. Заключение

В даной работе представлены результаты двух тематических исследований, посвященных применению разработанного нами параллельного алгоритма PSF [8] для аннотирования

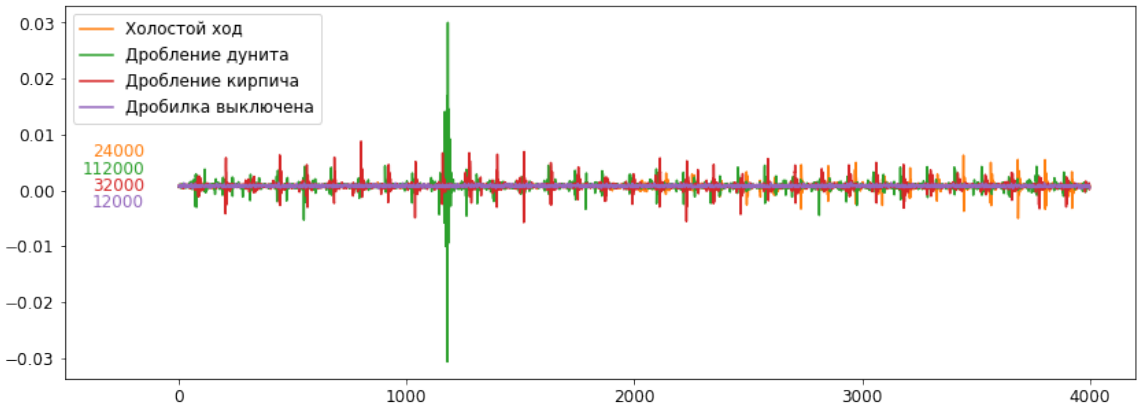


Рис. 5. Найденные сниппеты (числа, данные цветом, показывают индексы начала сниппетов)

сенсорных данных на графическом процессоре. Приведено краткое описание методов реализации разработанного алгоритма. Первое исследование связано с аннотированием показаний носимого виброакселерометра, закрепленного на человеке. В среднем точность классификации не ниже 74%. Наибольший показатель точности был достигнут для подъема по лестнице, наименьший – для глажки. Второе исследование связано с аннотированием данных виброаксерометра, установленного на малогабаритной дробильной установке. Записанные данные включают дробление дунита и кирпича (твердого и мягкого материалов соотвественно). Наибольшая точность аннотирования была достигнута для дробления дунита – 72%. Во всех исследованных случаях параллельный алгоритм [8] показал большую эффективность по сравнению с оригинальной последовательной версией [1].

Литература

1. Imani S., Madrid F., Ding W. et al. Matrix Profile XIII: Time Series Snippets: A New Primitive for Time Series Data Mining // Proceedings of the 9th IEEE International Conference on Big Knowledge (ICBK), Singapore, 2018. P. 382–389. DOI: 10.1109/ICBK.2018.00058.

2. Mueen A., Keogh E.J., Zhu Q. et al. Exact Discovery of Time Series Motifs // Proceedings of the 2009 SIAM International Conference on Data Mining (SDM). Sparks, Nevada, USA, 2009. P. 473–484. DOI: 10.1137/1.9781611972795.41.

3. Ye L., Keogh E.J. Time Series Shapelets: a New Primitive for Data Mining // Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009. P. 947–956. DOI: 10.1145/1557019.1557122.

4. Indyk P., Koudas N., Muthukrishnan S. Identifying Representative Trends in Massive Time Series Data Sets Using Sketches // Proceedings of the 26th International Conference on Very Large Data Bases (VLDB). Cairo, Egypt, 2000.

5. Gharghabi S., Imani S., Bagnall A.J. et al. An ultra-fast time series distance measure to allow data mining in more complex real-world deployments // Data Mining and Knowledge Discovery. 2020. Vol. 34. P. 1104–1135. DOI: 10.1007/s10618-020-00695-8.

6. Reiss A., Stricker D. Introducing a New Benchmarked Dataset for Activity Monitoring // Proceedings of the 16th International Symposium on Wearable Computers (ISWC). Newcastle, United Kingdom, 2012. P. 108-109. DOI: 10.1109/ISWC.2012.13.

7. Yeh C.M., Zhu Y., Ulanova L. et al. Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets // Proceedings of the IEEE 16th International Conference on Data Mining (ICDM). Barcelona, Spain, 2016. P. 1317–1322. DOI: 10.1109/ICDM.2016.0179.

8. Цымблер М.Л., Гоглачев А.И. Поиск типичных подпоследовательностей временного ряда на графическом процессоре // Вычислительные методы и программирование. 2021. 22, № 4. 344–359. DOI: 10.26089/NumMet.v22r423.