ТОПОЛОГИЧЕСКИЕ МЕТОДЫ АНАЛИЗА В СИСТЕМАХ ПОВЕДЕНЧЕСКОЙ АНАЛИТИКИ

Нашивочников Н.В.[1], Пустарнаков В.Ф.[2]

Цель статьи: *разработка методики применения методов анализа больших данных, основанных на топологических конструкциях, применительно к системам поведенческой аналитики для обеспечения корпоративной и киберфизической безопасности.*

Метод: *методика основана на алгебраической теории персистентных гомологий. Наряду с алгебраической топологией используются эмбедология (теория вложения Такенса-Мане) и теория метрических пространств.*

Полученный результат: *даются необходимые понятия алгебраической топологии, лежащие в основе анализа профилей поведения пользователя/сущности: симплициальный комплекс Виеториса-Рипса, фильтрация по множеству точек облака, группы гомологий, модули персистентности, топологические характеристики и зависимости. На первом этапе методики временные ряды, которые описывают изменяющееся во времени поведение пользователя/сущности, преобразуются в облако точек топологического пространства. Для указанного преобразования применяются методы теории вложения Такенса-Мане и алгоритм метода ложных соседей. На последующих этапах методики для базового и текущего облаков точек строятся топологические зависимости, диаграммы (персистентности, бар-кодов), характеризующие базовый и текущий профили поведения соответственно. На заключительном этапе выявляется отклонение текущего профиля поведения от базового. Для оценивания отклонения используются метрики Вассерштейна, Чебышева, узкого места и шкалирование на основе обобщенной функции желательности Харрингтона. Приводятся результаты практической апробации предложенной методики применения топологических алгоритмов к данным системы мониторинга работы пользователей корпоративной сети с информационными ресурсами.*

Ключевые слова: *аналитика поведения пользователей и сущностей, профиль поведения, вычислительная топология, персистентная гомология, временные ряды, эмбедология, кластеры, кибербезопасность.*

DOI: 10.21681/2311-3456-2021-2-26-36

1. Введение

В современной отрасли кибербезопасности наблюдается устойчивый интерес к системам поведенческой аналитики (User and Entity Behavior Analytics (UEBA)) [1,2] – новому классу решений безопасности, основанных на интеллектуальной обработке данных от учетных записей пользователей и объектов (устройств, приложений, сетей и т.д.) корпоративных и киберфизических систем [3]. Для выявления аномалий в поведении пользователей/сущностей, которые могут представлять собой инциденты безопасности, в UEBA наряду со статистическими методами находят применение методы расширенной аналитики[3], включая алгоритмы глубокого обучения [4]. В последние годы для поиска закономерностей и выявления аномалий в сложных данных большего объема заметное развитие получили топологические методы анализа (topology data analysis (TDA)) [5-16]. Появились первые работы, посвященные применению TDA в области кибербезопасности [17]. Результаты, представленные в [18,19], свидетельствуют о повышении результативности и оперативности визуального обнаружения подозрительных событий безопасности, получаемых от систем обнаружения вторжений [18] или систем управления информацией и событиями безопасности [19], благодаря применению TDA.

Характерной особенностью UEBA является построение модели типового поведения (базового профиля) пользователей/сущностей. При определенном отклонении поведения пользователя/сущности от базового профиля, UEBA фиксирует потенциальный инцидент безопасности. В отличие от [18,19], посвященных задаче визуального анализа событий безопасности с использованием топологической кластеризации на основе алгоритма Mapper [5,8], предметом настоящей работы является методика анализа профилей поведения пользователей/сущностей на основе теории персистентных гомологий [5, 7-9,12,13].

1 Нашивочников Николай Васильевич, CISSP, заместитель генерального директора – технический директор ООО «Газинформсервис», г. Санкт-Петербург, Россия. E-mail: cto@gaz-is.ru

2 Пустарнаков Валерий Фёдорович, кандидат физико-математических наук, первый заместитель генерального директора ООО «Газинформсервис», г. Санкт-Петербург, Россия. E-mail: pustarnakov-v@gaz-is.ru

3 Advanced Analytics // Gartner IT Glossary. URL: https://www.gartner.com/it-glossary/advanced-analytics/

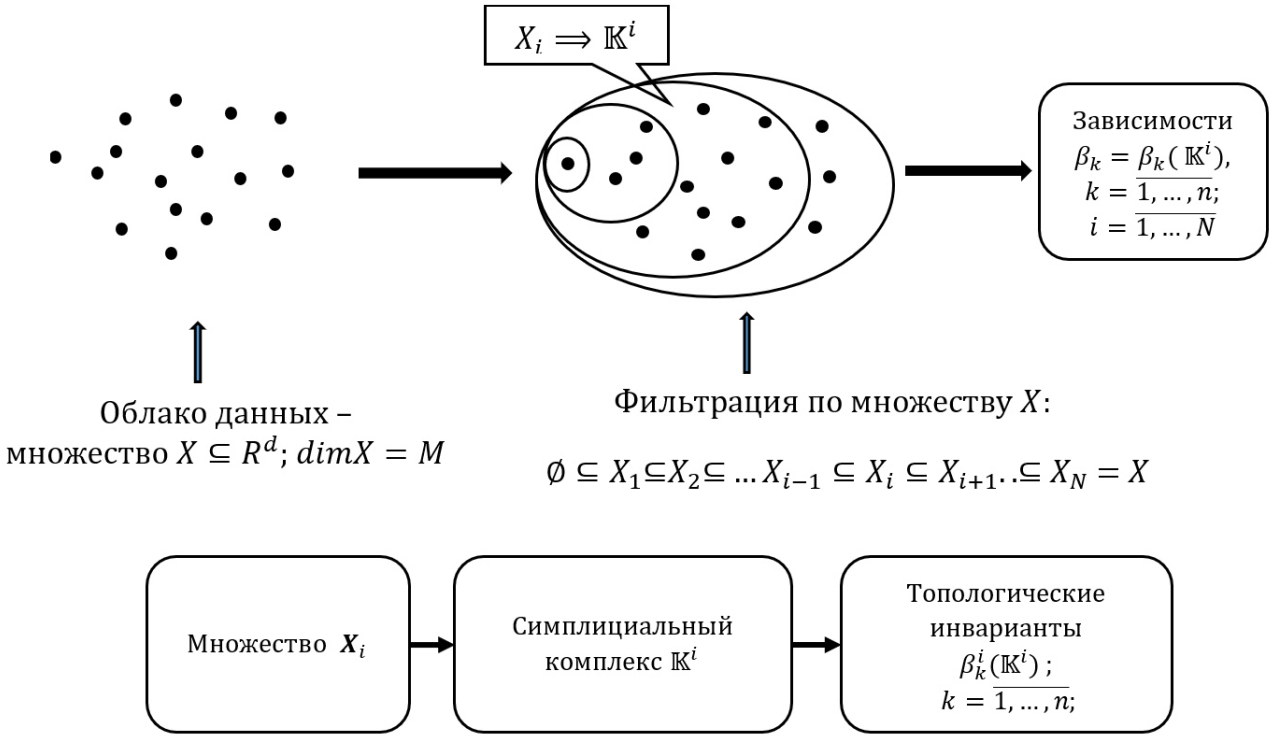

Рис.1. Общая схема TDA

2. Методика применения TDA в UEBA

При топологическом подходе исходными данными для построения и сравнения базового и текущего профилей пользователя/сущности выступают облака данных – неупорядоченный набор данных, не привязанный к какой-либо шкале измерений, аналогичной временной. Облака данных представляются в виде множеств точек[4] в некотором топологическом пространстве, к которому применяются процедуры TDA (рис.1). Представленные на рис.1 понятия и обозначения раскрываются ниже, по мере описания этапов предлагаемой методики применения процедур TDA в UEBA.

Этап 1. Преобразование временного ряда в облако точек

В UEBA основная часть исходных данных представляет собой временные ряды. Поэтому на первом этапе методики временной ряд преобразуется в облако точек без потери информации, т.е. подбирается такое топологическое пространство, элементами которого и будут являться элементы временного ряда. После определения топологического пространства, включающего в себя облако точек, становится возможным вычисление топологических инвариантов и их производных характеристик для выявления особенностей анализируемого временного ряда.

Предположим, что облако точек содержится в топологическом пространстве X, являющимся евклидовым пространством размерности d. Облако точек покрывается симплексами[5] следующим образом. Пусть P – облако точек в R^d:

$$P = \{v_i | \, i = \overline{1,2,\dots,N}\}, \quad (1)$$

где $v_i \in R^d, N \leq d$.

Введем в рассмотрение матрицу $D_E = \{D_E(v_i, v_j)\}_{i,j=1,1}^{N,N}$, элементы которой $D_E(v_i, v_j)$ вычисляются следующими образом:

$$D_E(v_i, v_j) = \|v_i - v_j\| \triangleq \sqrt{\langle v_i - v_j, v_i - v_j \rangle}, \quad (2)$$

где $\|\cdot\|$ — норма в евклидовом пространстве $R^d, \|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}, i, j = 1, \dots, N$.

Для каждого $v_i \in P$ определим множество $B_\lambda(v_i)$:

$$B_{\lambda/2}(x, v_i) = \{x | \, D_E(x, v_i) \leq \lambda/2, x \in R^d\}, \quad (3)$$

где $\lambda \in [0, U]$, $U = \max_{i,j} D_E(v_i, v_j)$.

Множество $B_{\lambda/2}(x, v_i)$ представляет собой замкнутый шар радиуса $\lambda/2$ в R^d.

Введем в рассмотрение симплициальный комплекс Виеториса-Рипса $\mathbb{K}(\lambda)$ [5,15]. Пусть задан параметр λ. Тогда говорят, что симплекс $\sigma^N(\lambda) = [v_0 v_1 \dots v_N] \in \mathbb{K}(\lambda)$ тогда и только тогда, когда для любых точек $v_{i_1}, v_{i_2} \in \sigma^N(\lambda)$ выполняется условие:

$$D_E(v_{i_1}, v_{i_2}) \leq \lambda, 1 \leq i_1, i_2 \leq N, \quad (4)$$

где $v_{i_1}, v_{i_2} \in P$.

Таким образом, для данного значения λ симплициальный комплекс $\mathbb{K}(\lambda)$ представляет собой множество симплексов таких, что для любых двух симплексов

4 Каждому элементу в облаке данных ставится в соответствие точка в соответствующем облаке

5 Эта процедура называется триангуляцией

$\sigma_{v^{(1)}}(\lambda)$, $\sigma_{v^{(2)}}(\lambda) \in \mathbb{K}(\lambda)$ выполняются следующие условия:

$$\sigma_{v^{(1)}}(\lambda) \cap \sigma_{v^{(2)}}(\lambda) \in \mathbb{K}(\lambda), \quad (5)$$

$$\sigma'(\lambda) \subset \sigma_{v^{(1)}}(\lambda) \vee \sigma_{v^{(2)}}(\lambda) \Rightarrow \sigma'(\lambda) \in \mathbb{K}(\lambda). \quad (6)$$

Симплициальный комплекс $\mathbb{K}(\lambda)$, состоящий из $(p+1)$ точек (триангулированный различными симплексами Виеториса-Рипса[6]), является p-мерным симплициальным комплексом. При этом размерность p комплекса $\mathbb{K}(\lambda)$ не превышает N–1 в случае, если исходное облако имеет N точек.

Далее, пусть временной ряд $K(t) = \{k_t, t = 1, 2, \dots, T\}$ отображает наблюдения за промежуток времени T. В общем случае этот ряд порождается некоторым априори неизвестным процессом (динамической системой). Согласно теории вложения Такенса-Мане [20] описание фазового пространства динамической системы можно получить, взяв вместо реальных переменных системы (которые в общем случае неизвестны) некоторые d–мерные векторы задержек, составленные из значений временного ряда $K(t)$ в последовательные моменты времени.

При выполнении условия $d \geq 2m + 1$, где m – размерность «настоящего» аттрактора, возможно реконструировать фазовое пространство (пространство состояний) контролируемой (наблюдаемой) системы, т.е. преобразовать временной ряд в облако точек в пространстве R^d. Такой процесс преобразования называется построением копии аттрактора в топологическом пространстве (в нашем случае это евклидовое пространство R^d) размерности $d \geq 2m+1$, где m – фрактальная размерность «настоящего» аттрактора. Алгоритм преобразования заключается в следующем [21-23]. Возьмем d последовательных отсчетов временного ряда $K(t)$. Далее, начиная с произвольного номера n, формируем последовательность $\{k_n, k_{n+1}, \dots, k_{n+d-1}\}$, где $d \geq 2m+1$. Элементы этой последовательности представляют собой компоненты d-мерного вектора и образуют точку в R^d. Следующая точка получается сдвигом нового d-мерного набора на величину $\tau > 0$, а итеративное продолжение этой процедуры даст последовательность точек в R^d — искомое облако данных. Теорема Такенса [20] гарантирует, что такое вложение сохранит свойства временного ряда с точностью до диффеоморфизмов[7]. Это означает, что для топологического вложения можно использовать любую непрерывную функцию, а сдвиг, который был использован, – самая простая функция из возможных.

Для определения параметров d и τ будем использовать алгоритм, базирующийся на методе ближайшего ложного соседа [24]. Метод основан на идее совпадения геометрических и топологических свойств исходного и восстановленного аттракторов, т.е. геометрические и топологические свойства, содержащиеся в исходном временном ряде, не теряются при преобразованиях.

Алгоритм метода ложных ближних соседей состоит в следующем:

1. Фиксируем размерность d. Пусть $x \in R^d$ – точка в R^d ,(x-вектор). Находим для каждой точки x_i ближайшего соседа x_j в d-мерном пространстве. Для этого вычисляем норму разности векторов x_i и x_j, т.е. $\|x_i - x_j\|$ и в качестве ближайшего соседа выбираем тот вектор, для которого эта норма минимальна, т.е.:

$$x_{j(n)} = \min_{x_j \in R^d} \|x_i - x_j)\|; \ \forall i, j = \overline{1, \dots d}. \quad (7)$$

2. Вычисляем норму разности векторов x_i, $x_{j(n)}$ $\|x_i - x_{j(n)})\|$.

3. После этого проводим одну итерацию и подсчитываем величину R_i:

$$R_i = \frac{\|x_{i+1} - x_{j+1(n)})\|}{\|x_i - x_{j(n)})\|}. \quad (8)$$

4. Если $R_i > R_{\text{П}}$, где $R_{\text{П}}$ – априори заданное пороговое значение, то такая точка является ложным ближним соседом. В результате подсчитывается количество таких ложных ближних соседей P.

5. Вычисляется отношение $\frac{P}{N}$ и алгоритм повторяется для $d=d+1$.

Здесь N– общее количество точек восстановленного пространства, а частное $\frac{P}{N}$ – оценка меры достоверности текущего значения d, поскольку определяется отношением количества ложных соседей (восстановленных точек) к их общему числу.

6. Итерационный процесс продолжается до тех пор, пока не выполнится условие:

$$\left|\frac{P}{N}\right| \leq \varepsilon, \quad (9)$$

где $\varepsilon \geq 0$ – априори заданное сколь угодно малое число. На практике, начиная с некоторой размерности d^0, отношение $\frac{P}{N}$ может стабилизироваться относительно весьма малого, но отличного от ε значения. Тогда в качестве искомой размерности принимается значение d^0. Таким образом, с использованием процедур этапа 1 временной ряд преобразуется в облако точек в топологическом (в нашем случае – евклидовом) пространстве R^d. Далее, на этапах 2-4 к облаку точек последовательно применяются процедуры TDA согласно схеме на рис.1.

Этап 2. Фильтрация по множеству X

Осуществляется фильтрация по множеству X:

$$\emptyset \subseteq X_1 \subseteq X_2 \subseteq \dots X_{i-1} \subseteq X_i \quad (10)$$

Строится система вложенных множеств так, как показано на рис.1. Пусть $|X| = M$ ($|X|$ – мощность множества, в данном случае $|X|$ – количество точек). Из набора перенумерованных данных берем первую точку (элемент набора данных) – $x_1 \in R^d$ и полагаем: $X_1 = \{x_1\}$; $|X_1| = 1$. Далее строим множество X_2. Для этого к элементу x_1 добавляем еще m элементов из облака данных. Тогда $|X_2| = |X_1| + m = 1 + m$. Процесс продолжается до тех пор, пока множество X_N совпадет с исходным множеством облака точек X. Выбор чисел N и шага фильтрации $|X_i \backslash X_{i-1}| < M$ связан с анализом природы порождающего временной ряд процессов.

6 В TDA могут использоваться и другие комплексы, например, комплексы Чеха, Вороного, Делоне, Альфа-комплекс. Выбор симплекса Виеториса-Рипса обусловлен возможностью его достаточно простой программной реализации

7 Непрерывных и дифференцируемых преобразований

Этап 3. Построение симплициального комплекса $\mathbb{K}^i$

Для каждого множества X_i для выбранной априори меры близости $\lambda>0$ с помощью алгоритма Вьеториса-Рипса, строится симплициальный комплекс $\mathbb{K}^i$.

Для комплекса $\mathbb{K}^i$ вычисляются группы гомологий:

$$H_k^i(\mathbb{K}^i) = Z_k^i(\mathbb{K}^i)/B_k^i(\mathbb{K}^i), k = \overline{1,\dots,n}, \quad (11)$$

где $Z_k^i(\mathbb{K}^i) = Ker\partial_k(\mathbb{K}^i); B_k^i = Im\partial_{k+1}(\mathbb{K}^i)$.

Топологические инварианты (числа Бетти) $\beta_k^i(\mathbb{K}^i)$ для множества X_i рассчитываются по формулам:

$$\beta_k^i(\mathbb{K}^i) = \mathrm{rank}H_k^i(\mathbb{K}^i) = \mathrm{rank}Z_k^i(\mathbb{K}^i) - \mathrm{rank}B_k^i(\mathbb{K}^i). \quad (12)$$

Так как R^d – векторное пространство, то:

$$\beta_k^i(\mathbb{K}^i) = dimZ_k^i(\mathbb{K}^i) - dimB_k^i(\mathbb{K}^i). \quad (13)$$

Этап 4. Построение топологических зависимостей и диаграмм

Строятся зависимости $\beta_k(i) = \{(i,\beta_k^i) \in \mathbb{R}^2 | i = \overline{1,\dots,N}; \beta_k^i \in \mathbb{R}\}$[8], бар-коды и диаграммы персистентности соответствующих групп гомологий как для анализируемого, так и для базового облаков данных[9].

Построение бар-кодов и диаграмм персистентности выполняется следующим образом [13].

Зафиксируем λ для симплициального комплекса $\mathbb{K}(\lambda)$. Пусть $\varphi_\lambda(\cdot)$ –вещественная непрерывная функция, определенная на симплексах, входящих в комплекс $\mathbb{K}(\lambda)$, и удовлетворяющая условию:

$$\varphi_\lambda(\sigma_q^N) \le \varphi_\lambda(\sigma_l^N) \quad (14)$$

для всех симплексов $\sigma_q^N, \sigma_l^N \in \mathbb{K}(\lambda); q,l = \overline{1,N}; q \ne l$.

Построим последовательность $\Phi(x)$:

$$\Phi(x) = \{\varphi_\lambda(\sigma^N) \le x, \sigma^N \in \mathbb{K}(\lambda); x \in \mathbb{R}\}. \quad (15)$$

Элементы последовательности $\Phi(x)$, очевидно, также являются симплициальными комплексами:

$$\Phi(x) \subseteq \Phi(y), \forall x \le y;\ x,y \in \mathbb{R}. \quad (16)$$

Тогда любой последовательности $\{x_1<x_2<\dots<x_n\}$ можно поставить в соответствие упорядоченную последовательность включений[10]:

$$\emptyset \subset \Phi(x_1) \subset \Phi(x_2) \subset \cdots \subset \Phi(x_n) \subset \mathbb{K}(\lambda). \quad (17)$$

Данная последовательность в свою очередь инициирует следующую последовательность преобразований групп гомологий [5]:

$$H_k(\Phi(x_1)) \xrightarrow{\varphi_{\lambda(1)}} H_k(\Phi(x_2)) \xrightarrow{\varphi_{\lambda(2)}} \dots \xrightarrow{\varphi_{\lambda(n-1)}} H_k(\Phi(x_n)). \quad (18)$$

Для векторных топологических пространств отображения:

$$\varphi_{\lambda(i,j)}(\cdot)\colon H_k(\Phi(x_i)) \to H_k(\Phi(x_j))\ \forall i,j \in \{1,\dots,l\}; \quad (19)$$

являются линейными.

Из свойств функтуриальности (отображение симплициальных комплексов индуцирует отображение на группах гомологий) [9] следует, что:

$$\varphi_{\lambda(l,i)} \circ \varphi_{\lambda(j,l)} = \varphi_{\lambda(i,j)}\ \forall\ i \le k \le j. \quad (20)$$

Модулем персистентности называют пару [9]:

$$\left\{H_k(\Phi(x_i))_{1\le i\le l}, \{\varphi_{\lambda(i,j)}\}_{1\le i\le l\le l}\right\}, \quad (21)$$

где для всех $i,j \in \{1,\dots,l\}$; $i \le j$.

В процессе преобразования групп гомологий одни группы появляются, а другие исчезают в модуле персистентности. Представляет практический интерес именно слежение за появлением и исчезновением нетривиальных групп гомологий в указанном модуле по мере прохождения фильтрации [9,12]. Обозначим их через b и d_e (от англ. birth и death – рождение и смерть соответственно).

Пусть B – множество промежутков $[b, d_e)$. Введем в рассмотрение интервальный модуль $I_{[b,d_e)}$ [12,16]. Интервальный модуль является модулем персистентности, для которого $\varphi_{\lambda(i)} = 1$, если $i \in [b,d_e) \subset B$ и $\varphi_{\lambda(i)} = 0$ – в противном случае. Эти интервалы легко точно закодировать, отметив, когда появляются и исчезают группы гомологий. Геометрически они отображаются бар-кодами.

Кроме бар-кодов в TDA используются диаграммы персистентности, которые представляют собой альтернативный графический способ представления бар-кодов.

Диаграммы персистентности формируются следующим образом [9]. Промежуток $[b, d_e)$ представляется точкой в расширенной плоскости $\mathbb{R}^2$, где $\mathbb{R} = \mathbb{R} \cup \{\infty\}$. Таким образом, диаграмма персистентности – это мультимножество[11], которое представляет собой объединение конечного мультимножества точек в $\mathbb{R}^2$ с мультимножеством точек на диагонали $\Delta = \{(x,y) \in \mathbb{R}^2 | x = y\}$, где каждая точка диагонали имеет бесконечную кратность. Диагональные точки включаются в диаграмму, чтобы иметь возможность их сравнивать при изучении взаимно однозначного соответствия между принадлежащими им точкам. Это предполагает, что диаграммы должны быть множествами с одинаковой мощностью. Такое сравнение чрезвычайно полезно в UEBA.

Пусть $q \in [1,\infty]$, $d(\cdot,\cdot)$ – метрика в L^q для $q \in [1,\infty]$[12], $\xi(\cdot)$[12] $\xi(\cdot)$ – биекция диаграммы A на E. Тогда мера близости между диаграммами A, E (метрика Вассерштейна) определяется следующим образом [16]:

$$W_l[d](A,E) = \inf_{\xi(\cdot):A\to E} [\sum_{a\in A} d[a,\xi(a)]^q]^{\frac{1}{q}}, \quad (22)$$

для $q \in [1,\infty]$;

$$W_\infty[d](A,E) = \inf_{\xi(\cdot):A\to E} \sup_{a\in A} d[a,\xi(a)], \quad (23)$$

при $q=\infty$[13].

8 Иногда эти зависимости нормируются, чтобы использовать для их сравнения метрики близости вероятностных распределений

9 Индексом (0) вверху будем обозначать базовый профиль

10 Данная процедура называется фильтрацией множества (симплициального комплекса) $\mathbb{K}(\lambda)$ по φ_λ

11 Мультимножество - это множество элементов, в которое каждый элемент может входить больше одного раза

12 L^q-пространство измеримых функций, модуль которых в q-й степени имеет конечный интеграл Лебега

13 Одной из наиболее часто используемых метрик близости при сравнении диаграмм персистентности является метрика узкого места W_∞ [d]

DOI: 10.21681/2311-3456-2021-2-26-36

Этап 5. Выявление отклонения от базового профиля

1. Зависимости $\beta_k = \beta_k(i)$ сглаживаются с помощью кубической сплайн-интерполяции:

$$\beta_k = S_{k(3)}(x), \quad (24)$$

где $S_{k(3)}(\cdot)$ – кубический сплайн, $x \in \mathbb{R}; 1 \le x \le N$. При этом узлы интерполяции сплайна $S_{k(3)}(\cdot)$ должны совпадать с узлами сплайна для базового профиля $-S^0_{k(3)}(\cdot)$.

Так как зависимости $\beta_k = S_{k(3)}(x)$ являются непрерывными функциями, то для оценки их отклонения от базового профиля можно использовать метрику Чебышева:

$$\rho_k\left(S_{k(3)}(x), S^{(0)}_{k(3)}(x)\right) = \max_x \left|S_{k(3)}(x) - S^{(0)}_{k(3)}(x)\right|, \quad (25)$$
$$k = \overline{0, \dots, n}.$$

Если облако точек есть множество в $\mathbb{R}^2$, то числа Бетти, начиная с β_2, будут равны нулю. Это означает, что метрика Чебышева вычисляется для $k = 0; 1$, т.е. $\rho_0(\cdot,\cdot)$ и $\rho_1(\cdot,\cdot)$. Если зависимости $S_{k(3)}(x)$; $S^{(0)}_{k(3)}(x)$ – нормированы и трактуются как распределения, то для их сравнения можно использовать дивергенцию Кульбака-Лейблера [25][14]:

$$KL_k\left(S_{k(3)} \parallel S^{(0)}_{k(3)}\right) = -\int S_{k(3)}(x) log \frac{S^{(0)}_{k(3)}}{S_{k(3)}(x)} dx, \quad (26)$$
$$k = \overline{0, \dots, n}.$$

2. Для каждого комплекса $\mathbb{K}^i$ подмножества $X_i, i = \overline{1, \dots, N}$ рассчитываются бар-коды и строятся диаграммы персистентности.

Отклонения от базового профиля оценивается с помощью мер близости Вассерштейна (22) или узкого места (23).

В результате реализации пунктов 1 и 2 на этапе 5 формируется кортеж:

$$\langle \rho_k\left(S_{k(3)}(x), S^{(0)}_{k(3)}(x)\right), \{W_{l,i}[d](A_i, A^0_i)\}^N_{i=1}, \quad (27)$$
$$\{W_{\infty,i}[d](A_i, A^0_i)\}^N_{i=1}\rangle,$$

где $A_i = A(X_i)$ – диаграмма персистентности для множеств X_i фильтра $\{X_i\}^N_{i=1}$; $A^0_i = A^0(X_i)$ – диаграмма персистентности i-го базового профиля; $k = \overline{0, \dots, n}$.

В практических приложениях порой достаточно вычислить метрики Вассерштейна и узкого места для $A_N = A(X_N), A^0_N = A^0(X_N)$. Тогда выражение (27) запишется в виде:

$$\langle \rho_k\left(S_{k(3)}(x), S^{(0)}_{k(3)}(x)\right), W_{l,N}[d](A_N, A^0_N), \quad (28)$$
$$W_{\infty,N}[d](A_N, A^0_N)\rangle.$$

Кортеж (27) или (28) в достаточной степени характеризует поведение пользователя/сущности, основываясь только на доступных данных о его активности в информационном среде. Рассмотрим, как оценить поведение пользователя/сущности, базируясь на этих данных и результатах проведенного по описанной выше методики анализа.

Для простоты ограничимся метрикой узкого места. Сформируем вектор $V \in \mathbb{R}^{n+1}$:

$$V = \left\{\rho_k\left(S_{k(3)}(x), S^{(0)}_{k(3)}(x)\right), W_{\infty,N}[d](A_N, A^0_N)\right\}^T \quad (29)$$
$$k = \overline{0, \dots, n}.$$

Поскольку V является вектором основных показателей, описывающих поведение пользователя/сущности[15], то оценивание «характера» поведения представляет собой задачу многокритериального выбора [26,27].

Вынося за рамки настоящей работы исследование возможных подходов к разрешению проблемы выбора, используем для оценки поведения наблюдаемого объекта такой показатель как «обобщенная функция желательности Харрингтона» [28]. Эта «обобщенная функция» появилась в результате наблюдений за реальными решениями, принимаемыми экспериментаторами, и обладает такими полезными свойствами, как гладкость и монотонность.

В предлагаемой методике для оценки поведения пользователя/сущности используем процедуру формирования решающего правила, которое формулируется на основе обобщенной функции желательности Харрингтона и оценок отдельных показателей:

$$\rho_k\left(S_{k(3)}(x), S^{(0)}_{k(3)}(x)\right); W_{\infty,N}[d](A_N, A^0_N); k = \overline{0, \dots, n}, \quad (30)$$

полученных в результате применения изложенной выше методики.

Для расчета обобщенной оценки берется одна из логистических функций Харрингтона («кривая желательности»):

$$\gamma(x) = exp[-exp(-y(x)], \quad (31)$$

где $y(x)$ – кодированные значения частных показателей (скалярная величина), x – скалярная переменная.

Функция (31) была выведена эмпирическим путем. Ось абсцисс называется шкалой частных показателей. В нашем случае это отклонения:

$$\rho_k\left(S_{k(3)}(x), S^{(0)}_{k(3)}(x)\right); W_{\infty,N}[d](A_N, A^0_N); k = \overline{0, \dots, n}. \quad (32)$$

Отклонения кодируются. Ось ординат – шкала желательности. Функция $\gamma(x)$ имеет два участка насыщения $(\gamma \to 0; \gamma \to 1)$ и практически линейный участок (от $\gamma = 0{,}2$ до $\gamma = 0{,}63$). Промежуток эффективных значений на шкале частных показателей (оси абсцисс) – [-2; +5]. Шкала желательности делится в интервале от 0 до 1 на пять диапазонов (табл.1).

Таблица 1

Шкала желательности

Номер диапазона	Наименование градации (лингвистические значения)	Числовые интервалы
1	Очень плохо	0–0,2
2	Плохо	0,2–0,37
3	Удовлетворительно	0,37–0,63
4	Хорошо	0,63–0,8
5	Очень хорошо	0,8–1

14 Дивергенция Кульбака-Лейблера мера несимметричная, т.е. $KL_k\left(S_{k(3)} \parallel S^{(0)}_{k(3)}\right) \neq KL_k\left(S^{(0)}_{k(3)} \parallel S_{k(3)}\right), \forall k = \overline{0, \dots, n}$

15 Показатели отклонения поведения пользователя/сущности от базового профиля вычисляются с помощью топологических инвариантов. Являются безразмерными величинами

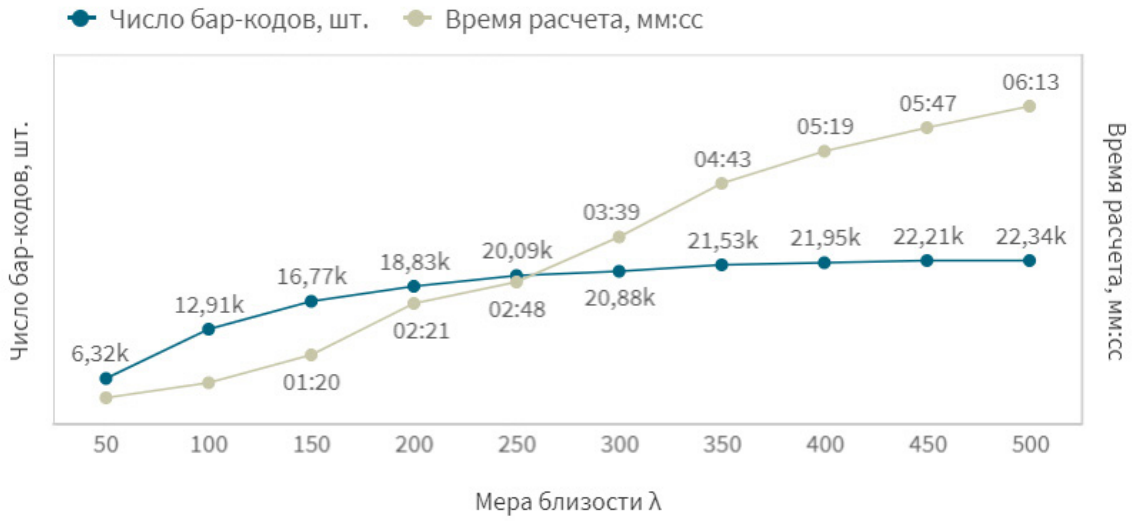

Рис.2. Выбор меры близости λ

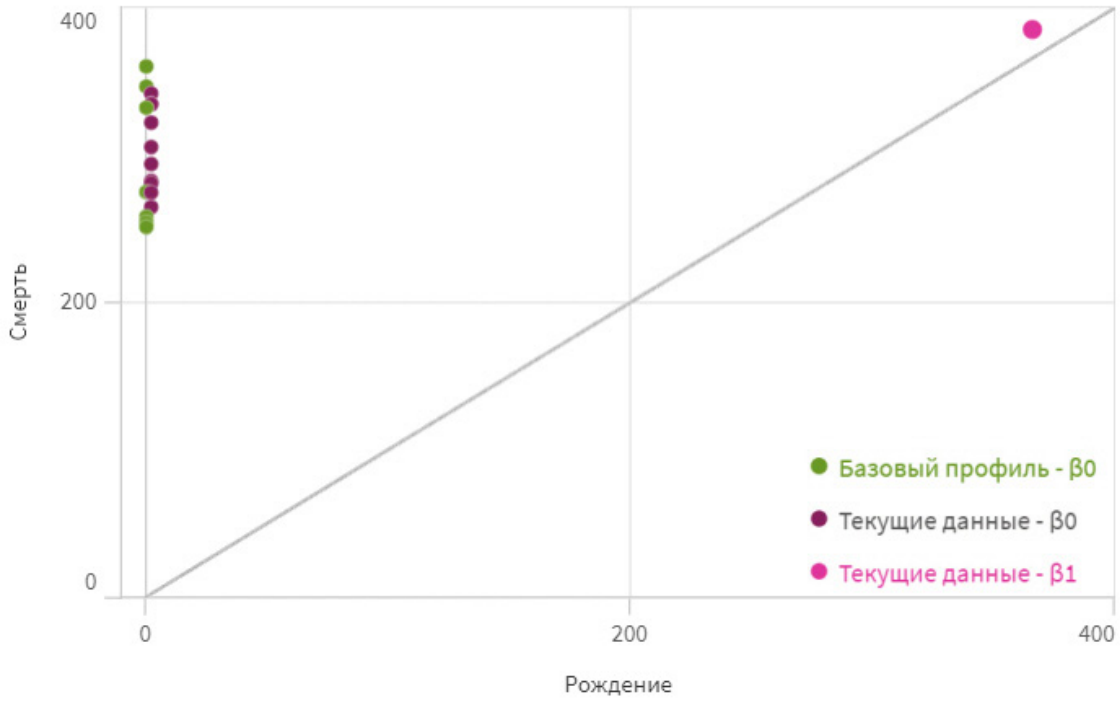

Рис.3. Диаграмма персистентности для параметра «Поиск работы»

Обобщенный показатель желательности $\sum$ вычисляется следующим образом [26]:

$$\Sigma = \sqrt[r]{\prod_{i=1}^{r} \gamma_i}, \quad (33)$$

где r – количество используемых показателей (в нашем случае $n+1$).

Причем корень r-й степени «сглаживает» возникающие флуктуации, являясь в некотором роде фильтром.

3. Экспериментальные результаты

В связи с пандемией коронавириса повысилась актуальность задачи анализа поведения пользователей корпоративной сети, работающих дистанционно. Для решения указанной задачи использовались данные об активности работы с информационными ресурсами корпоративной сети и сети Интернет. Данные за февраль 2020 г. (до перехода на удаленный режим работы) использовались для формирования базового профиля.

DOI: 10.21681/2311-3456-2021-2-26-36

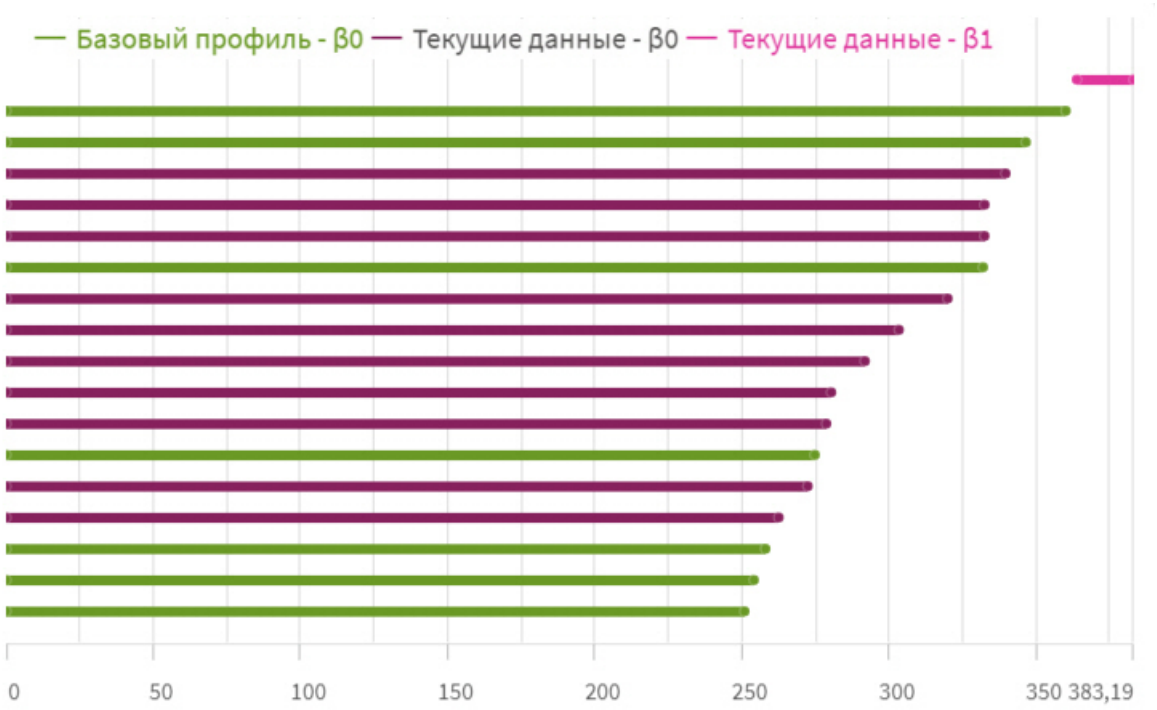

Рис.4. Бар-коды для параметра «Поиск работы»

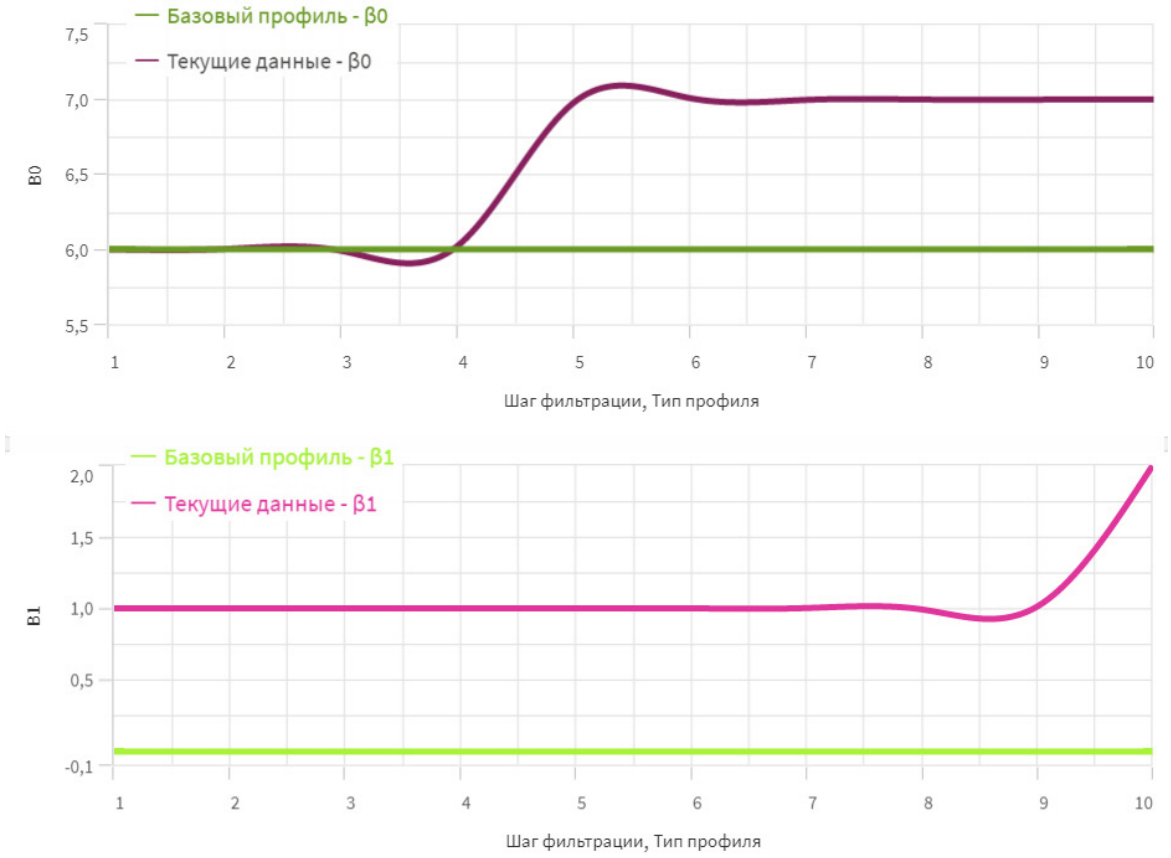

Рис.5. Профили персистентности для параметра «Поиск работы»

Анализировалось поведение в марте 2020 г. (после перехода на режим удаленной работы). Данные были закодированы и преобразованы в облака точек[16] согласно описанной выше методике. Для каждой активности размерность пространства получилась равной двум.

Выбор меры близости в комплексе Витриеса-Рипса λ=400 осуществлялся автоматически, путем контроля числа новых бар-кодов при увеличении λ. Если при увеличении λ количество новых бар-кодов оставалось меньше 5% от их общего числа, то расчет останавливался. На рис.2 приведены показатели расчета параметра λ от 50 до 500 с шагом в 50 условных единиц при 17 268 точек в облаке точек[17].

Параметры фильтрации также рассчитывались автоматически для каждого числа Бетти. Фильтр определялся отдельно и таким образом, чтобы было отобрано не более 50 стойких топологических особенностей. В будущем для их выбора планируется использовать методы машинного обучения.

Для примера на рис.3 и рис.4 приведены диаграмма персистентности и бар-коды соответственно. На них отражены данные об активности обращений конкретного пользователя корпоративной сети к информационным ресурсам, связанным с поиском работы.

На рис.5 представлены построенные профили персистентности $-S_{k(3)}(x), S^{(0)}_{k(3)}(x), k = 0,1.$

Далее были рассчитаны величины отклонений от базового профиля – метрики Чебышева для профилей персистентности $\rho_0\left(S_{0(3)}(x), S^{(0)}_{0(3)}(x)\right) = 1.09348$, $\rho_1\left(S_{1(3)}(x), S^{(0)}_{1(3)}(x)\right) = 1.994793$, метрика узкого места $W_\infty[d] = 39.3772$.

Закодировав показатели $\rho_0(\cdot,\cdot)$, $\rho_1(\cdot,\cdot), W_\infty[d]$ согласно предлагаемой методике, был рассчитан обобщенный показатель желательности $\sum$=0.8547. Согласно табл.1, показатель соответствует градации «очень хорошо», что позволяет сделать вывод: поведение пользователя в части поиска работы при переходе на удаленный режим существенно не изменилось. Этот же вывод можно сделать, анализируя изменения профили персистентности. Так, на рис.5 не наблюдается появления дополнительных «дыр» в топологической структуре («дыра» появляется на уровне топологического шума) при фактическом их отсутствии в базовом профиле (число Бетти – β_1), а компонента связанности (число Бетти – β_0) изменяется в большую сторону, что говорит о повышении уровня группирования (кластеризации) данных. Таким образом, можно сделать вывод о том, что поведение объекта исследования в «среднем» изменилось не существенно.

Следует подчеркнуть, что этот вывод был сделан только по активности «поиск работы». По другим активностям, связанным непосредственно со служебной деятельностью выбранного пользователя, ситуация была также «хорошей».

4. Заключение

Предложенная в работе методика впервые раскрывает возможности топологического подхода для такой актуальной задачи как разработка систем поведенческой аналитики в обеспечении кибербезопасности. Перспективными направлениями дальнейшей работы представляются:

- совершенствование методик поведенческой аналитики путем совместного применения алгоритмов TDA и машинного обучения;
- апробация TDA в решении аналитических задач для различных типов систем обеспечения безопасности, в частности, в системах операционного мониторинга и анализа киберфизических систем различных классов [3].

Литература

1. Матвеев А. Обзор рынка систем поведенческого анализа – User and Entity Behavioral Analytics (UBA/UEBA). URL: https://www.anti-malware.ru/analytics/Market_Analysis/user-and-entity-behavioral-analytics-ubaueba.
2. Sadowski G., Litan A., Bussa T., Phillips T. Market Guide for User and Entity Behavior Analytics. Published: 23 April 2018. ID: G00349450. Gartner. 2018.
3. Нашивочников Н.В. Проблемные вопросы применения аналитических средств безопасности киберфизических систем предприятий ТЭК / Нашивочников Н.В., Большаков А.А., Николашин Ю.А., Лукашин А.А. // Вопросы кибербезопасности №5 (33). 2019. С. 26-33.
4. Chalapathy R., Chawla S. Deep Learning for Anomaly Detection: A Survey. arXiv:1901.03407 [cs.LG]. 2019. URL: https://arxiv.org/pdf/1901.03407.pdf.
5. Carlsson G. Topology and data // Bull. of the Amer. Mathem. Soc. 2009. Vol. 46(2), P. 255–308.
6. Offroy V, Duponchel L, Topological data analysis: A promising big data exploration tool in biology, analytical chemistry and physical chemistry // Analytica chimica acta. 2016. vol. 910. P. 1-11. https://doi.org/10.1016/j.aca.2015.12.037.
7. Wasserman L. Topological Data Analysis. arXiv:1609.08227v1 [stat.ME]. 2016. URL: https://arxiv.org/pdf/1609.08227.pdf.
8. Chazal F., Bertrand M. An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists. arXiv: 1710.04019 [math.ST]. 2017. URL: https://arxiv.org/pdf/1710.04019.pdf.
9. Chazal F., de Silva V., Glisse M., Oudot S. The Structure and Stability of Persistence Modules. Springer International Publishing. 2016. P. 120. DOI: 10.1007/978-3-319-42545-0.
10. Huntsman S., Palladino J., Robinson M. Topology in cyber research. arXiv:2008.03299 [math.AT]. 2020. URL: https://arxiv.org/pdf/2008.03299.pdf.

16 Оказалось, что даже эти неполные и зашумленные данные содержали полезную информацию

17 Для триангуляции облака данных и расчета топологических инвариантов использовалось программное обеспечение с открытым кодом Ripser [15]

11. Brüel-Gabrielsson R., Nelson B., Dwaraknath A., Skraba P, Guibas L., Carlsson G. A Topology Layer for Machine Learning. arXiv:1905.12200v2 [cs.LG]. 2020. URL: https://arxiv.org/pdf/1905.12200.pdf.
12. Otter N., Porter M.A., Tillmann U. Grindrod P., Harrington H.A. A roadmap for the computation of persistent homology // EPJ Data Science. 6, 17. 2017. https://doi.org/10.1140/epjds/s13688-017-0109-5.
13. Kerber M., Schreiber H. Barcodes of Towers and a Streaming Algorithm for Persistent Homology // Discrete & Computational Geometry volume. 2019. v.61, P. 852-879. https://doi.org/10.1007/s00454-018-0030-0.
14. Love E.R., Filippenko B., Maroulas V., Carlsson G. Topological Deep Learning. arXiv:2101.05778 [cs.LG]. 2021. URL: https://arxiv.org/pdf/2101.05778.pdf.
15. Bauer U. Ripser: efficient computation of Vietoris-Rips persistence barcodes. arXiv:1908.02518 [math.AT]. 2019. URL: https://arxiv.org/pdf/1908.02518.pdf
16. Arjovsky M., Chintala S., Bottou L. Wasserstein Generative Adversarial Networks // Proceedings of the 34th International Conference on Machine Learning, PMLR. 2017. P. 214-223.
17. Chow Y. Application of Data Analytics to Cyber Forensic Data // Worcester Polytechnic Institute: BS Thesis , 2016. P. 100.
18. Coudriau M., et al., Topological analysis and visualisation of network monitoring data: Darknet case study // IEEE International Workshop on Information Forensics and Security (WIFS). 2016. P. 1-6.
19. Trevor J. Bihl, Robert J. Gutierrez, Kenneth W. Bauer, Bradley C. Boehmke, Cade Saie. Topological Data Analysis for Enhancing Embedded Analytics for Enterprise Cyber Log Analysis and Forensics // Proceedings of the 53rd Hawaii International Conference on System Sciences, 2020. P. 1937-1946. DOI: 10.24251/HICSS.2020.238.
20. Макаренко Н.Г. Эмбедология и нейропрогноз. Часть 1. – М. МИФИ. 2003. 188 с.
21. Фомичев А.В. Элементы теории бифуркаций и динамических систем. Часть II. – М. МФТИ. 2019, 50 с.
22. Барышева Е.Н., Никишов В.Н. Модели оценки финансовых показателей с учетом их стохастичности и хаотичности // Вестник СамГУ. 2012. № 4 (95). С. 115-126.
23. Рюэль Д. Случайность и хаос. – М. Издательство «Регулярная и хаотическая динамика». 2001. 192 с.
24. Krakovská A., Mezeiová K., Budáčová N. Use of False Nearest Neighbours for Selecting Variables and Embedding Parameters for State Space Reconstruction // Journal of Complex Systems, 2015. P. 1-12. https://doi.org/10.1155/2015/932750.
25. Постовалов С.Н. Статистический анализ данных, моделирование и исследование вероятностных закономерностей. Компьютерный подход / Лемешко Б.Ю., Лемешко С.Б., Постовалов С.Н., Чимитова Е.В. – Новосибирск. Изд-во НГТУ. 2011. 888 с.
26. Подиновский В.В., Ногин В.Д. Парето-оптимальные решения многокритериальных задач. – М. Физматлит. 2007. 256 с.
27. Орлов А.И. Методы принятия управленческих решений. – М. КНОРУС. 2018. 286 с.
28. Пичкалев А. В. Применение кривой желательности Харрингтона для сравнительного анализа автоматизированных систем контроля // Вестник КГТУ. – Красноярск: КГТУ. №1(1). 1997. С. 128-132.

TOPOLOGICAL METHODS OF ANALYSIS IN BEHAVIORAL ANALYTICS SYSTEMS

Nashivochnikov N.[18], Pustarnakov V.F.[19]

Purpose of the article: *development of a methodology for the application of methods for analyzing big data based on topological constructions in relation to behavioral analytics systems to ensure corporate and cyber-physical security.*

Method: *the technique is based on the algebraic theory of persistent homology. Along with algebraic topology, embedology (Takens-Mane embedding theory) and the theory of metric spaces are used.*

Result: *the necessary concepts of algebraic topology are given, which underlie the analysis of user / entity behavior profiles: Vietoris-Rips simplicial complex, filtering by a set of cloud points, homology groups, persistence modules, topological characteristics and dependencies. At the first stage of the technique, the time series that describe the time-varying behavior of the user / entity are transformed into a cloud of points in the topological space. For this transformation, the methods of the Takens-Mane embedding theory and the algorithm of the method of false neighbors are used. At the subsequent stages of the methodology for the base and current point clouds, topological dependencies, diagrams (persistence, bar codes) characterizing the base and current behavior profiles, respectively, are built. At the final stage, the deviation of the current behavior profile from the baseline is revealed. To estimate the deviation, the Wasserstein, Chebyshev, bottleneck metrics and scaling based on the generalized Harrington desirability function are used. The results of practical testing of the proposed method of applying topological algorithms to the data of the monitoring system for the work of corporate network users with information resources are presented.*

Keywords: *user and entity behavioral analytics, behavior profile, computational topology, persistent homology, time series, embedology, clusters, cybersecurity.*

References

1. Matveev A. Obzor ry`nka sistem povedencheskogo analiza – User and Entity Behavioral Analytics (UBA/UEBA). URL: https://www.anti-malware.ru/analytics/Market_Analysis/user-and-entity-behavioral-analytics-ubaueba.
2. Sadowski G., Litan A., Bussa T., Phillips T. Market Guide for User and Entity Behavior Analytics. Published: 23 April 2018. ID: G00349450. Gartner. 2018.
3. Nashivochnikov N.V. Problemny`e voprosy` primeneniya analiticheskix sredstv bezopasnosti kiberfizicheskix sistem predpriyatij TE`K / Nashivochnikov N.V., Bol`shakov A.A., Nikolashin Yu.A., Lukashin A.A. // Voprosy` kiberbezopasnosti №5 (33). 2019. S. 26-33.
4. Chalapathy R., Chawla S. Deep Learning for Anomaly Detection: A Survey. arXiv:1901.03407 [cs.LG]. 2019. URL: https://arxiv.org/pdf/1901.03407.pdf.
5. Carlsson G. Topology and data // Bull. of the Amer. Mathem. Soc. 2009. Vol. 46(2), P. 255–308.
6. Offroy V, Duponchel L, Topological data analysis: A promising big data exploration tool in biology, analytical chemistry and physical chemistry // Analytica chimica acta. 2016. vol. 910. P. 1-11. https://doi.org/10.1016/j.aca.2015.12.037.
7. Wasserman L. Topological Data Analysis. arXiv:1609.08227v1 [stat.ME]. 2016. URL: https://arxiv.org/pdf/1609.08227.pdf.
8. Chazal F., Bertrand M. An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists. arXiv: 1710.04019 [math.ST]. 2017. URL: https://arxiv.org/pdf/1710.04019.pdf.
9. Chazal F., de Silva V., Glisse M., Oudot S. The Structure and Stability of Persistence Modules. Springer International Publishing. 2016. P. 120. DOI: 10.1007/978-3-319-42545-0.
10. Huntsman S., Palladino J., Robinson M. Topology in cyber research. arXiv:2008.03299 [math.AT]. 2020. URL: https://arxiv.org/pdf/2008.03299.pdf.
11. Brüel-Gabrielsson R., Nelson B., Dwaraknath A., Skraba P, Guibas L., Carlsson G. A Topology Layer for Machine Learning. arXiv:1905.12200v2 [cs.LG]. 2020. URL: https://arxiv.org/pdf/1905.12200.pdf.
12. Otter N., Porter M.A., Tillmann U. Grindrod P., Harrington H.A. A roadmap for the computation of persistent homology // EPJ Data Science. 6, 17. 2017. https://doi.org/10.1140/epjds/s13688-017-0109-5.
13. Kerber M., Schreiber H. Barcodes of Towers and a Streaming Algorithm for Persistent Homology // Discrete & Computational Geometry volume. 2019. v.61, P. 852-879. https://doi.org/10.1007/s00454-018-0030-0.
14. Love E.R., Filippenko B., Maroulas V., Carlsson G. Topological Deep Learning. arXiv:2101.05778 [cs.LG]. 2021. URL: https://arxiv.org/pdf/2101.05778.pdf.
15. Bauer U. Ripser: efficient computation of Vietoris-Rips persistence barcodes. arXiv:1908.02518 [math.AT]. 2019. URL: https://arxiv.org/pdf/1908.02518.pdf

18 Nikolay Nashivochnikov, CISSP, Deputy General Director - Technical Director, Gazinformservice LLC, St. Petersburg, Russia. E-mail: cto@gaz-is.ru

19 Valery Pustarnakov, Ph. D., First Deputy General Director, Gazinformservice LLC, St. Petersburg, Russia. E-mail: pustarnakov v@gaz-is.ru

16. Arjovsky M., Chintala S., Bottou L. Wasserstein Generative Adversarial Networks // Proceedings of the 34th International Conference on Machine Learning, PMLR. 2017. P. 214-223.
17. Chow Y. Application of Data Analytics to Cyber Forensic Data // Worcester Polytechnic Institute: BS Thesis , 2016. P. 100.
18. Coudriau M., et al., Topological analysis and visualisation of network monitoring data: Darknet case study // IEEE International Workshop on Information Forensics and Security (WIFS). 2016. P. 1-6.
19. Trevor J. Bihl, Robert J. Gutierrez, Kenneth W. Bauer, Bradley C. Boehmke, Cade Saie. Topological Data Analysis for Enhancing Embedded Analytics for Enterprise Cyber Log Analysis and Forensics // Proceedings of the 53rd Hawaii International Conference on System Sciences, 2020. P. 1937-1946. DOI: 10.24251/HICSS.2020.238.
20. Makarenko N.G. E`mbedologiya i nejroprognoz. Chast` 1. – M. MIFI. 2003. 188 s.
21. Fomichev A.V. E`lementy` teorii bifurkacij i dinamicheskix sistem. Chast` II. – M. MFTI. 2019, 50 s.
22. Bary`sheva E.N., Nikishov V.N. Modeli ocenki finansovy`x pokazatelej s uchetom ix stoxastichnosti i xaotichnosti // Vestnik SamGU. 2012. № 4 (95). S. 115 126.
23. Ryue`l` D. Sluchajnost` i xaos. – M. Izdatel`stvo «Regulyarnaya i xaoticheskaya dinamika». 2001. 192 s.
24. Krakovská A., Mezeiová K., Budáčová N. Use of False Nearest Neighbours for Selecting Variables and Embedding Parameters for State Space Reconstruction // Journal of Complex Systems, 2015. P. 1-12. https://doi.org/10.1155/2015/932750.
25. Postovalov S.N. Statisticheskij analiz danny`x, modelirovanie i issledovanie veroyatnostny`x zakonomernostej. Komp`yuterny`j podxod / Lemeshko B.Yu., Lemeshko S.B., Postovalov S.N., Chimitova E.V. – Novosibirsk. Izd-vo NGTU. 2011. 888 s.
26. Podinovskij V.V., Nogin V.D. Pareto optimal`ny`e resheniya mnogokriterial`ny`x zadach. – M. Fizmatlit. 2007. 256 s.
27. Orlov A.I. Metody` prinyatiya upravlencheskix reshenij. – M. KNORUS. 2018. 286 s.
28. Pichkalev A. V. Primenenie krivoj zhelatel`nosti Xarringtona dlya sravnitel`nogo analiza avtomatizirovanny`x sistem kontrolya // Vestnik KGTU. – Krasnoyarsk: KGTU. №1(1). 1997. S. 128 - 132.