УДК 004.056.5 **DOI** 10.15622/sp.54.3

Е.С. Новикова, И.Н. Муренин

МЕТОДИКА ВИЗУАЛЬНОГО АНАЛИЗА МАРШРУТОВ СОТРУДНИКОВ ОРГАНИЗАЦИИ ДЛЯ ОБНАРУЖЕНИЯ АНОМАЛИЙ

Новикова Е. С., Муренин И. Н. **Методика визуального анализа маршрутов сотрудников организации для обнаружения аномалий.**

Аннотация. Обнаружение аномалий в перемещениях сотрудников является важной задачей, которая связана с обеспечением киберфизической безопасности предприятий, включая критические инфраструктуры. В работе представлен подход к анализу перемещений сотрудников критической инфраструктуры, отличающийся сочетанием алгоритмов интеллектуального анализа данных и интерактивных методик визуализации. Он включает в себя два этапа — определение групп сотрудников с похожим поведением и обнаружение аномалий. Группировка пользователей по их поведению осуществляется с помощью самоорганизующихся карт Кохонена; для отображения пространственно-временных шаблонов поведения используется разработанная авторами модель визуализации BandView. Для обнаружения аномалий в поведении сотрудников предлагается механизм оценки значений пространственно-временных атрибутов движения. Отображение отклонений осуществляется с помощью тепловой карты, позволяющей аналитику с легкостью определить зону и интервал времени с подозрительной активностью. Подход апробирован на наборе данных, предоставленном в рамках конкурса VASTMiniChallenge-2 2016, который описывает перемещения сотрудников внутри здания организации.

Ключевые слова: выявление аномалий в траекториях, визуальная аналитика, паттерны поведения, оценка отклонений в поведении, тепловые карты.

1. Введение. В настоящее время данные, описывающие перемещения движущихся точек в пространстве, являются наиболее распространенным типом пространственных данных, и их анализ имеет множество практических применений. Они могут быть получены с помощью датчиков мобильных телефонов, автомобилей, считывателей карт доступа, видеокамер, расположенных в местах наблюдений за объектами. Значения атрибутов движения объектов могут быть получены как напрямую от наблюдаемых объектов, так и от специализированных сенсоров, регистрирующих их появление. Обычно такие наборы данных называют траекториями. Исследование траекторий помогает определить шаблоны поведения объектов, выявить ограничения, существующие в исследуемой среде, например, правила или политики безопасности, регулирующие права доступа к определенным зонам, или социально значимые объекты, например, банкоматы, кафе, аптеки и так далее [1-3]. Другим важным применением анализа траекторий является создание шаблонов поведения объектов для последующего обнаружения возможных аномалий в их маршрутах. В этом случае анализ траекторий может использоваться для обеспечения безопасно-

сти движения воздушного, наземного и водного транспорта путем мониторинга местоположения, траектории, скорости движения транспортного средства и обнаружения неожиданных препятствий на маршруте. Мониторинг перемещений сотрудников критических инфраструктур обеспечивает контроль соблюдения мер и политик безопасности, установленных на предприятии [4].

В настоящее время на рынке программных продуктов представлено большое количество решений, осуществляющих мониторинг перемещений сотрудников внутри организации, а также отслеживающих их действия на рабочих местах, контролируя список запущенных процессов, набираемый текст, делая аудио- и видеозапись [5-8]. Основной задачей таких систем является контроль активности сотрудников организации на рабочих местах, заключающийся в расчете рабочего времени, оценке динамики опозданий, продуктивности работы. Вместе с тем, согласно [9], анализ поведения пользователей должен позволять строить паттерны поведения, на основе которых в дальнейшем возможно выявление различных аномалий, которые могут свидетельствовать о таких угрозах, как внутренний нарушитель, финансовые мошенничества и целенаправленные атаки.

В работе представлен подход к выявлению пространственно-временных шаблонов перемещений персонала организации и аномалий в них, отличающийся сочетанием автоматических методик анализа данных и графического представления данных. Множество интерактивных моделей визуализации не только осуществляет графическое представление выявленных шаблонов в перемещениях сотрудников и потенциально-аномальных ситуаций, но и позволяет контролировать результаты применения автоматических методов анализа данных. Паттерны поведения отображаются с помощью разработанной авторами модели визуализации BandView, отражающей последовательность контролируемых зон, посещаемых сотрудником. Отклонения в перемещениях сотрудников отображаются в виде тепловой карты. Для уменьшения возможных шумов на тепловой карте предлагается механизм оценки отклонений, позволяющий сфокусироваться на отдельных подозрительных выбросах. Оценка аномалий осуществляется на основе пространственно-временных атрибутов отклонений, таких как место, продолжительность пребывания и время посещения. Сценарий использования методики и ее эффективность продемонстрирована на множестве тестовых данных, предложенных в рамках конкурса VAST Challenge 2016 [10].

Основной вклад авторов заключается в разработке подхода к визуальному анализу движений сотрудников организации, оперирующему разнородными данными, включающими должность сотрудников,

расположение их рабочих мест и временные атрибуты их перемещений для обеспечения киберфизической безопасности.

2. Методики анализа и визуализации траекторий. В случае отсутствия каких-либо априорных данных о типичном поведении (движении) объектов для анализа траекторий чаще всего применяются алгоритмы кластеризации данных. Полученные кластеры используются в дальнейшем для описания модели нормального поведения объектов, на основе которой производится последующее обнаружение аномалий. Для поиска кластеров применяются подходы, основанные на центроидах и методах, иерархических моделях, а также на основе плотности распределения объектов в пространстве признаков [11]. Автоматические методы могут обнаруживать интересные шаблоны поведения и аномалии, но для понимания и объяснения полученных результатов необходимо использовать дополнительные методики графического представления данных [12, 13].

Метод нейронной кластеризации, также известный как самоорганизующаяся карта Кохонена (self-organizing maps, SOM-карта), сочетает многомерную кластеризацию и методику проецирования признаков объектов из многомерного пространства в двумерное с сохранением расстояния между ними, обеспечивая при этом визуализацию полученного распределения объектов по кластерам. Шрек и другие применили нейронную сеть SOM для исследования траекторий и предложили инструмент визуального анализа данных, позволяющий контролировать результаты кластеризации, полученные с помощью сети SOM [14]. Основной акцент в работе сделан на анализ пространственных атрибутов траекторий. В [15] с помощью сети SOM исследуются как временные, так и пространственные атрибуты, однако их анализ выполняется путем выделения всех временных атрибутов для каждого географического местоположения или всех пространственных атрибутов для определенного интервала времени.

В общем случае существующие модели графического представления траекторий могут быть разделены на 3 группы: статические или интерактивные географические карты, часто дополняемые глифами, отвечающими за отображение атрибутов перемещений (1); пространственно-временные кубы (2) и вложенные графики с временной шкалой (3) [16].

Географические карты являются наиболее очевидным способом графического представления данных, описывающим изменения в местоположении объектов с течением времени. Траектории объектов или группы объектов отображаются в виде линий [17-19]. Атрибуты перемещений, такие как время, скорость, тип движущегося объекта кодируются цветом линии или отображаются с помощью специальных

глифов. В случаях, когда отдельные траектории не значимы, используются карты потоков [18, 19]. Карты потоков представляют собой модели визуализации, акцентирующие внимание на направлении движения, начальной и конечной точках маршрутов. Количественные характеристики потоков обычно отображаются с помощью цвета линии потока и ее насыщенности. Интересная модификация карт потоков используется в [19] для анализа миграций популяции. Географические регионы, выступающие в качестве источника или пункта назначения миграционных потоков, располагаются по окружности в виде сегментов кольца. Потоки отображаются в виде линий, соединяющих соответствующие сегменты кольца. Следует отметить, что для всех моделей визуализации траекторий на основе географических карт характерен один общий недостаток — они не могут отображать пространственные и временные атрибуты движения одновременно.

Пространственно-временной куб — это трехмерная модель визуализации, разработанная для одновременного представления пространственных и временных характеристик движущихся объектов. В этой модели точки траекторий отображаются в трехмерное пространство, где одна из осей, чаще всего вертикальная, обозначает время. В [20] представлено расширение пространственно-временного куба, названное *стеной траекторий* (*trajectory wall*). В этой модели визуализации вертикальная ось трехмерного куба используется для графического представления некоторого подмножества траекторий, обладающих схожими пространственными характеристиками. Так как они могут быть упорядочены по времени в третьем измерении, такая модель может рассматриваться в качестве трехмерного пространственно-временного куба, где абсолютное время преобразовано в порядок следования траекторий. Как и многие трехмерные модели визуализации, трехмерные пространственно-временные кубы могут быть неэффективны из-за большого нагромождения траекторий и недостатка пространства для отображения.

Вложенный график с временной осью представляет собой традиционный линейный график, одна ось которого обозначает время. По оси Y откладываются значения пространственных атрибутов точек траектории. Поскольку на графике обычно отображается сразу несколько характеристик траектории или несколько траекторий, то график называется *вложенным*. Каждый маршрут представляется в виде кривой или полосы [21-23]. Полоса (площадь под кривой) может быть разделена на сегменты, окрашенные в соответствии со значениями атрибутов движения объекта.

В современных системах контроля доступа анализ перемещений сотрудников выполняется путем построения маршрутов их перемещений и тепловых карт движения, осуществляется расчет пройденных дистанций и контроль соблюдения временного регламента перемещений на основе заданных графиков посещения контролируемых зон [5, 6, 8]. В некоторых случаях возможны выявления фактов обмана, связанных с одномоментной отметкой нескольких человек или отметкой присутствия за коллег [5]. Задача формирования паттернов движения сотрудников в общем случае не решается, выявление аномалий формируется в основном на основе правил, построенных в соответствии с описанием должностных инструкций. Кроме того, отсутствуют механизмы визуального анализа данных, процесс исследования инцидентов безопасности заключается в работе с исходными данными, представленными в табличном виде или в формате видео.

Представленная методика анализа траекторий сотрудников наиболее близка к методике, представленной в [14], поскольку в ней также применяются нейронные сети SOM для выявления траекторий, обладающих схожими характеристиками. Однако в отличие от [14], в работе оцениваются как пространственные, так и временные атрибуты маршрутов сотрудников. Графическое представление сети SOM усилено использованием специального глифа, дающего краткую характеристику объектов, принадлежащих одному кластеру. Кроме того, в работе представлен механизм оценки обнаруженных аномалий, учитывающий периодичность их появления и ранжирующий их по значимости.

3. Подход к обнаружению аномалий в перемещениях пользователей. Разработанная авторами методика анализа перемещений сотрудников внутри организации позволяет ответить на следующие вопросы:

1) Если ли группы сотрудников с похожим поведением?

2) Есть ли какие-то особенности в перемещениях сотрудников, зависящие от времени, то есть имеет ли место периодичность в перемещениях в зависимости от роли сотрудника?

3) Каков общий шаблон перемещений сотрудников, принадлежащих к одной группе? Как он меняется в зависимости от дня недели? Как это соотносится с положением работника в организации?

4) Есть ли какие-либо значимые отклонения в перемещениях сотрудников?

5) Каков характер аномалий, то есть как часто, где и когда они происходят?

Последовательно отвечая на данные вопросы, аналитик формирует в первую очередь общее понимание существующих шаблонов

перемещений, а затем фокусируется на деталях, описывающих возможные аномалии. Таким образом, авторы считают, что эти вопросы определяют основные этапы анализа траекторий и в целом согласуются с мантрой поиска информации, сформулированной Б. Шнейдерманом: «*общее представление данных → масштабирование и фильтрация → детали по требованию*» [24]. Модели визуализации и лежащие в их основе методики анализа данных, применяемые в предложенном подходе, поддерживают описанную схему аналитического процесса.

Ключевыми элементами методики являются нейронные сети SOM, использующиеся для формирования групп сотрудников, имеющих похожие шаблоны перемещений, и тепловые карты, которые применяются для обнаружения периодов аномального поведения. Они дополняются двумя моделями визуализации — графом контролируемых зон и вложенной моделью визуализацией *BandView*, отражающей последовательность посещенных зон и длительность пребывания в них.

Методика анализа состоит из следующих основных этапов:

1) этап подготовки данных, преобразующий записи считывателей контроля доступа в формат, необходимый для их последующей обработки;

2) формирование групп пользователей с одинаковым поведением и их отображение с помощью нейронных сетей Кохонена;

3) определение периодичности поведения пользователей одной группы путем выявления дней недели с одинаковым поведением;

4) вычисление и формирование графического представления паттернов поведения внутри одной группы или для заданного дня недели;

5) выявление аномалий в траекториях путем статистической оценки отклонений от паттерна траектории и их графическое представление с помощью тепловой карты;

6) детальный анализ аномального участка маршрута сотрудника с помощью модели визуализации *BandView.*

На рисунке 1 представлен макет основного окна прототипа программного обеспечения, реализующего предложенную методику. Первая SOM-карта «Сотрудники» (A) осуществляет разбиение персонала на группы с похожими траекториями перемещений; вторая SOM-карта «Дни» оценивает периодичность в траекториях сотрудников, принадлежащих одной группе. Панель (C) имеет две вкладки, предназначенные для отображения шаблонов маршрутов и аномалий — Pattern View и Anomaly View соответственно. В обоих случаях используется модель визуализации BandView.

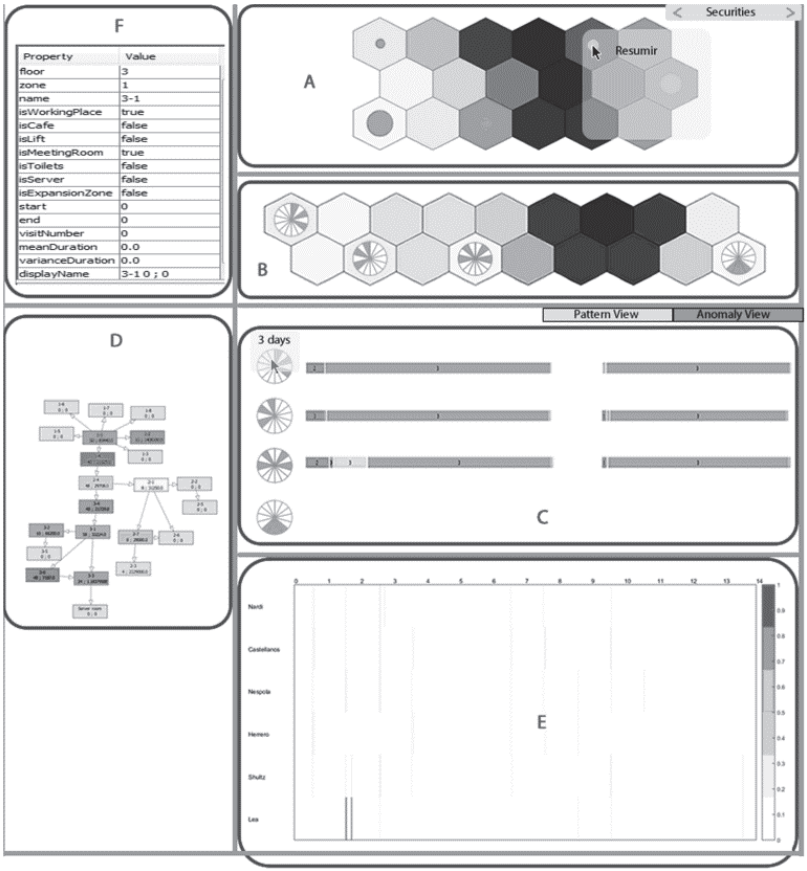


Рис. 1. Макет основного окна программного обеспечения, реализующего предложенную методику

Граф контролируемых зон (D) отражает маршрут выбранного сотрудника или группы сотрудников. Тепловая карта (E) демонстрирует отклонения в поведении для групп сотрудников или групп дней с похожими шаблонами перемещений. Свойства объектов, представленных различными графическими элементами моделей визуализации, отображаются с помощью таблицы свойств (F). Все модели графического представления данных интерактивны и связаны между собой. Выбирая мышкой элемент любой модели визуализации, аналитик обновляет информацию на остальных моделях, связанных с текущей. Например, при выборе элемента карты SOM «Сотрудники» происходит обновление остальных компонент графического интерфейса: таб-

лица свойств (F) наполняется подробной информацией о выбранной группе, SOM-карта «Дни» (B) отображает кластеры дней с похожими траекториями перемещений для выбранной группы, пространственно-временные шаблоны маршрутов для каждой группы дней представляются с помощью графа (D) и панели (C). Отклонения в траекториях внутри выбранной группы отображаются на тепловой карте (E). Выбор элемента на SOM-карте «Дни» позволяет сфокусироваться на определенном подмножестве дней, обновляя соответствующим образом компоненты (C)-(F). Следует отметить, что тепловая карта E непосредственно связана с вкладкой Anomaly View: выбор любой области тепловой карты приводит к обновлению модели BandView, расположенной на вкладке AnomalyView. Таким образом, SOM-карты могут рассматриваться как элементы графической фильтрации; а тепловая карта — как навигационная панель для выбора наиболее интересных для аналитика отклонений.

В следующих разделах описаны этапы предварительной обработки данных, а также приводятся подробные пояснения к предложенным методикам визуализации и анализа данных.

4. Этап подготовки исходных данных. В представленном подходе исходными данными являются журналы, формируемые датчиками контроля доступа и содержащие информацию только о времени посещения контролируемой зоны определенным сотрудником. Особенностью подхода является работа с журналами датчиков контроля доступа, и результат применения методики не зависит от технологии, используемой для реализации считывателей системы контроля доступа; это могут быть считыватели на основе rfid/nfc меток [6] или Bluetooth-маяков Beacon [5]. Важным моментом является возможность извлечения из логов системы контроля доступа следующей информации о нахождении сотрудника: время посещения контролируемой зоны и идентификатор данной зоны. Данные о должностях сотрудников и расположении контролируемых зон позволяют более точно описать паттерны передвижения, добавив к ним семантическую составляющую, например, указав, что в данной зоне находится рабочее место сотрудника.

Отличительная особенность логов, формируемых датчиками контроля доступа в том, что регистрация сотрудников происходит непосредственно в момент их появления в контролируемой зоне, вследствие чего записи в журнале носят нерегулярный характер и интервал между ними для определенного сотрудника может варьироваться от нескольких секунд до нескольких часов. Кроме того, некоторые сотрудники могут совершать множество перемещений по зданию организации в соответствии с их должностными обязанностями, в то

время как другие значительно реже покидают свои рабочие места. Таким образом, логи датчиков можно рассматривать как неравноотстоящие временные ряды переменной длины.

Одним из способов предварительной обработки неравноотстоящих временных рядов для их дальнейшего анализа является их дискретизация, позволяющая преобразовать их в векторы конечной длины.

Пусть $E = \{e_i\}_{i=1}^{n}$ — множество сотрудников, $Z = \{z_j\}_{j=1}^{m}$ — множество контролируемых зон, $T = \{t_k : t_i < t_j; i < j\}_{k=1}^{p}$ — упорядоченное множество временных меток, соответственно, записи из журнала считывателей контролируемых зон имеют вид (e_i, z_j, t_k). Общий интервал времени, представленный первой и последней записью журнала, обозначим как T_0, а множество записей обозначим как $LOGS = \{(e_i, z_j, t_k)\}, i = 1 \div n,\ j = 1 \div m, k = 1 \div p.$ Интервал времени T_0 разбивается на последовательность одинаковых временных интервалов $\Delta t,$ то есть:

$$T_0 = \{\Delta t_l : \Delta t_i = \Delta t_j; \Delta t_i = [t_i; t_{i+1}); \Delta t_{i+1} = [t_{i+1}; t_{i+2}); i \neq j; i, j \leq l; \}_{l=1}^{r}.$$

Для каждого интервала времени Δt_l и для каждого сотрудника e_i вычисляется число посещений $n_{z_j}^{\Delta t_l}$ и длительность пребывания $\Delta t_{z_j}^{\Delta t_l}$ в каждой контролируемой зоне z_j. Таким образом, множество LOGS можно представить в виде множества упорядоченных во времени пар $LOGS = \{(n_{z_j}^{\Delta t_l}; \Delta t_{z_j}^{\Delta t_l})\}_{e_i}, i = 1 \div n,\ j = 1 \div m, l = 1 \div r,$ вычисленных для каждого сотрудника $e_i, i = 1 \div n$. Это множество упорядоченных пар формирует множество атрибутов траектории пользователя, позволяя совместить пространственно-временные характеристики траектории сотрудников организации. Атрибуты вектора упорядочиваются сначала по наблюдаемому временному интервалу, а затем по идентификатору контролируемой зоны.

Длительность интервала времени Δt_l по умолчанию составляет 4 часа — эксперименты показали, что этого достаточно для того, чтобы обнаружить даже незначительные временные отклонения длительностью в 1 минуту.

5. Модель визуализации на основе нейронной сети SOM. К основным методам визуальной кластеризации следует отнести методы

главных компонент, многомерного шкалирования, метод t-SNE и другие. Данные алгоритмы осуществляют проекцию исходного многомерного пространства в пространство меньшей размерности, чаще всего двумерное, для последующего построения графического представления данных с помощью графиков рассеивания. Другим примером визуальной кластеризации является методика RadViz, которая также осуществляет проекцию многомерного пространства в двумерное, в основе которой лежит физическая метафора: каждый объект соединяется с n координатными узлами, обозначающими атрибуты объекта, n пружинами, жесткость которых зависит от значения соответствующих координат. Данные методики обладают общим свойством — снижая размерность признакового пространства объектов, они не снижают объем исходной выборки. Таким образом, они подходят как средство контроля результатов применения автоматических методов кластеризации и определения исходного числа кластеров. Однако для выявления некоторого шаблона данных, описываемого, например, центроидом кластера, они не подходят. По этой причине для обнаружения групп сотрудников с похожим поведением и индивидуальных шаблонов поведения для отдельных сотрудников используется нейронная сеть SOM (и ее графическое представление — SOM-карта). Ее основная задача — это разведывательный анализ данных при отсутствии информации о структуре исходных данных. Данный тип нейронной сети использует обучение без учителя для кластеризации многомерных данных [25]. Сеть SOM состоит из узлов или нейронов, с соответствующими весовыми векторами, размерность которых определяется исходными данными. В ходе итерационного обучения входные векторы сравниваются с весовыми векторами каждого нейрона, а веса нейрона, наиболее соответствующие входному вектору, и веса соседних нейронов регулируются так, чтобы быть ближе к входному вектору. Это обеспечивает возможность нейронной сети сохранять отношения между соседними нейронами, что означает, что кластеры, ассоциированные с узлами сети, расположенными рядом друг с другом, имеют большую степень сходства, чем кластеры, нейроны которых расположены далеко друг от друга. Одной из главных проблем сетей SOM является необходимость в непрерывной обработке данных без пропусков отдельных значений для каждого атрибута. Однако в предложенной методике этап предварительной обработки данных гарантирует получение векторов с явно определенными значениями для каждого атрибута.

Нейронные сети SOM могут рассматриваться в качестве методики визуализации данных, так как они обеспечивают графическое отображение многомерных данных в двумерное пространство. Для

представления структуры сети SOM часто используется U-матрица [26]. Она отражает структуру данных с помощью отображения среднего расстояния между весовыми векторами соседних нейронов. В предложенном подходе используется шестиугольная сетка, поэтому рассматривается вектор с 6 соседними узлами. Чем темнее цвет узла — тем дальше он находится от своих соседей. Узлы, окрашенные цветами с похожей интенсивностью, схожи друг с другом. Узлы, содержащие центры кластеров, отмечены глифом в форме круга, размер которого пропорционален количеству объектов в кластере.

В предложенном подходе нейронная сеть SOM используется дважды. Цель SOM-карты «Сотрудники» заключается в выявлении групп сотрудников с похожим поведением. Вектор признаков сотрудника описывает его активность на протяжении всего периода времени работы датчиков контроля доступа, в результате однократные или редкие отклонения в перемещениях не влияют на результат кластеризации. Это позволяет предположить, что SOM отображает различия в траекториях сотрудников в соответствии с особенностями их ролей в организации.

Уровень детализации анализа траекторий сотрудников пропорционален количеству узлов на SOM-карте. Выбор карты большего размера приводит к увеличению общего числа отображаемых кластеров. В таком случае ячейки сети SOM, соответствующие работникам с похожим поведением, будут сосредоточены в регионе карты, раскрашенном схожими оттенками. Эксперименты показали, что разница в поведении сотрудников, принадлежащих к соседним узлам, незначительна и объясняется, как правило, небольшими отклонениями в продолжительности пребывания в пределах контролируемой зоны, в то время как число посещенных зон и даже последовательности посещений сохраняются. Такие различия во времени редко превышают 5-10 минут, таким образом, есть возможность использовать одну общую пространственно-временную траекторию для всех сотрудников, принадлежащих к одному региону SOM. Для объединения таких сотрудников в один кластер рекомендуется настраивать размер карты SOM примерно равным их количеству. Однако разрабатываемый программный инструмент предусматривает возможность задавать различные размеры SOM-карты, чтобы исследовать возможные варианты группировки сотрудников.

SOM-карта для сотрудников на рисунке 2 содержит результаты кластеризации сотрудников одной должности. Из него следует, что данное множество сотрудников можно разбить на 5 групп, исходя из подобия их маршрутов в течение рабочего дня. Одна из них довольно многочисленна (группа №1), в то время как остальные состоят из одного-двух человек. Кроме того, траектории движения двух групп сотрудников

(группы №4 и №5), расположенные в правом верхнем углы SOM-карты, сильно отличаются, так как отделяются друг от друга множеством узлов, окрашенных в темный цвет. Дальнейший анализ их траекторий позволяет сделать вывод, что основные различия в их поведении объясняются расположением рабочих мест и должностными обязанностями.

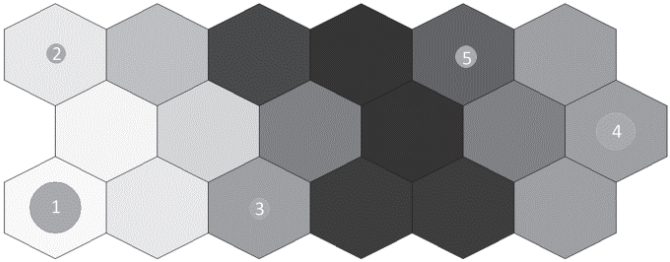


Рис. 2. SOM-карта «Сотрудники»

Авторы полагают, что сотрудники могут иметь обязанности, зависящие от дня недели, например, проведение еженедельного инструктажа, психологических тренингов и так далее. Подобные обязанности могут вызывать периодические изменения в перемещениях сотрудников. SOM-карта «Дни» позволяет выявить группы дней с похожими шаблонами перемещений сотрудников, принадлежащих к выбранному аналитиком кластеру, и определить таким образом периодичность в их маршрутах. Для построения данной SOM-карты используются центроиды кластеров SOM-карты «Сотрудники», преобразованные в векторы, соответствующие рабочим дням сотрудников, путем разбиения исходного вектора на векторы меньшей длины. Результат кластеризации нейронной сети также отображается с помощью U-матрицы, однако узлы SOM-карты дополнены специально разработанными глифами WeekCircle, отражающими распределение дней одного по дням недели. При проектировании глифа авторы учли, что во многих организациях деятельность сотрудников зависит от того, является ли неделя четной или нет. Глиф WeekCircle может быть использован для отображения шаблонов перемещений с периодической активностью в рамках одной или двух недель. В зависимости от заданной детализации он разделяется на 7 или 14 секторов, соответствующих дням недели. Модель глифа для двухнедельного периода перемещений представлена на рисунке 3. Правая половина глифа демонстрирует дни четной недели, а левая — нечетной. Понедельники обозначены верхними секторами, а выходные — нижними. На рисунке 3 изображены два глифа, которые отражают, какие дни недели включает группа дней. Левый глиф демонстрирует, что кластер дней состоит из понедельни-

ков, сред и пятниц четной недели, это означает, что выбранный сотрудник или группа сотрудников имеет определенные должностные обязанности, выполняемые каждые понедельник, среду и пятницу. Правый глиф показывает, что в группу дней вошли только выходные дни. Следует отметить, что SOM-карта «Дни» позволяет обнаружить дни с аномальным поведением, если эти аномалии имеют достаточно продолжительный характер, например, длятся более часа, поскольку дни с подобной активностью формируют отдельный кластер, расположенный по соседству с остальными.

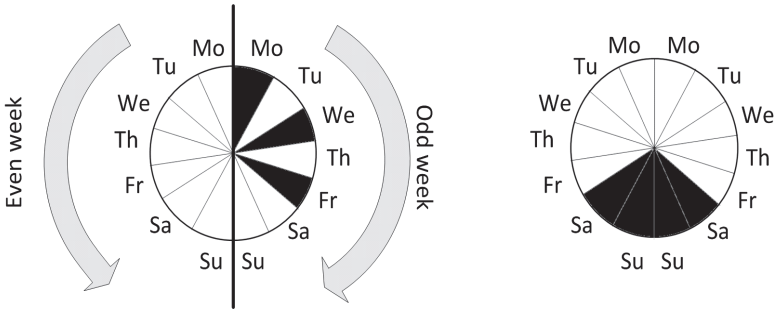


Рис. 3. Глиф WeekCircle, демонстрирующий разбиение множества дней на группы в соответствии с днем недели

6. Модель визуализации траекторий сотрудников BandView. Модель визуализации BandView предназначена для выявления связи между пространственными и временными атрибутами перемещений. Она представляет собой вложенную столбиковую диаграмму, горизонтальная ось которой обозначает время. Маршрут работника представлен с помощью последовательности блоков-сегментов, каждый из которых соответствует временному интервалу, в течение которого работник находится в данной зоне. Цвет сегмента используется для кодирования атрибутов зоны. В настоящее время цветовая схема кодирования контролируемых зон построена следующим образом: каждому этажу соответствует определенный цвет, а цвета для зон, расположенных на этом этаже создаются путем изменения яркости соответствующего цвета: чем больше значение идентификатора контролируемой зоны — тем темнее цвет соответствующего сегмента. На рисунке 4 представлены маршруты сотрудников отдела, выполняющих обслуживание здания в течение одного дня. Из него следует, что сотрудники данного отдела работают в три смены, выходя в ночное, дневное и вечернее время.

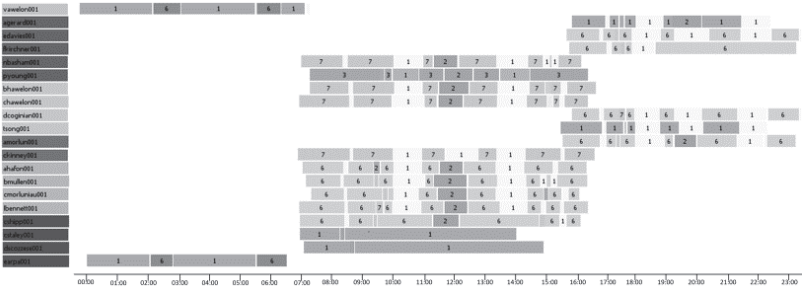


Рис. 4. Модель визуализации маршрутов сотрудников BandView

Очевидно, что продолжительность пребывания в зонах сильно варьируется — от нескольких секунд до нескольких часов, что приводит к тому, что длины сегментов полос BandView могут быть или очень маленькими и практически незаметными, или такими длинными, что для их просмотра требуется полоса прокрутки. Для того чтобы обеспечить аналитику возможность просматривать как короткие, так и длинные сегменты, реализован механизм масштабирования, построенный на нелинейном преобразовании шкалы времени, увеличивающем короткие интервалы времени и сокращающем длительные. Каждый сегмент модели может быть выбран с помощью мышки, подробная информация о нем, такая как продолжительность пребывания сотрудника в данной зоне, временная отметка, атрибуты зоны и так далее, отображаются в окне свойств.

Модель BandView используется для отображения шаблонов перемещений и сырых данных при исследовании отклонений в маршруте сотрудника. Следует отметить, что модель BandView позволяет легко обнаружить, где и когда произошла аномалия, как долго она длилась, благодаря возможности визуально сравнивать маршрут сотрудника с маршрутами его коллег.

Для работы с исходными данными реализован гибкий механизм фильтрации, позволяющий строить сложные логические выражения, используя всевозможные атрибуты перемещений: идентификатор сотрудника, должность, офис, продолжительность пребывания в зоне, идентификатор зоны, этаж и так далее.

7. Граф контролируемых зон. Методики визуализации на основе графов широко используются для представления пространственных атрибутов траекторий. В предложенном подходе вершины графа представляют контролируемые зоны, смежные зоны соединены ребрами. Зоны, посещенные сотрудником, выделены цветом в соответствии с цветовой схемой, используемой в BandView. Зоны, которые данный сотрудник не посещал, отмечены серым цветом. Граф контролируемых зон позволяет установить связь между посещенными сотрудником зонами и их атрибутами: наличие кафе, конференц-

зала, лестницы или лифта, расположение рабочего места сотрудника и его коллег и так далее. Исследование этих атрибутов позволяет объяснить причины посещения контролируемой зоны сотрудником. Для этого атрибуты контролируемых зон также кодируются цветом, и аналитик может выбирать режимы просмотра графа исходя из атрибутов контролируемых зон или их расположения внутри здания.

В предложенном подходе граф также используется для оценки статистических данных о перемещениях сотрудников: аналитику предоставляется информация о средней продолжительности пребывания в контролируемой зоне в течение дня и ее дисперсии, среднем числе посещений в день и общем количестве посещений на протяжении всего анализируемого интервала времени. Эти данные выводятся в таблице свойств объекта F и могут быть отражены в соответствующих вершинах графа. Кроме того, аналитик имеет возможность устанавливать размер вершины в зависимости от числа посещений или среднего времени пребывания в зоне. Авторы считают, что такая возможность способствует обнаружению зон, где локализованы рабочие места сотрудников, и зон, которые они посещают редко.

Так, например, на рисунке 5 изображен граф посещаемых зон, размер вершин которых зависит от количества посещений в течение всего рассматриваемого периода времени. Можно ясно увидеть, что сотрудник проводит большую часть времени в зоне 1-7, где расположено его рабочее место, а зону 3-4 посетил всего один раз. Поскольку в этой зоне находится лифт третьего этажа, можно предположить, что он посетил ее случайно, нажав не на ту кнопку лифта.

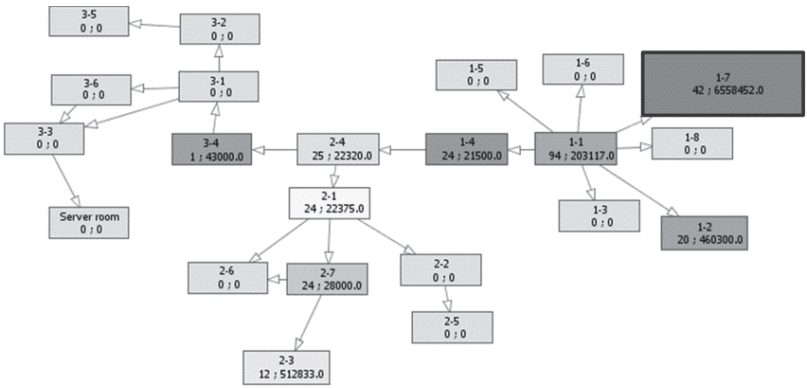


Рис. 5. Граф посещения контролируемых зон

Данные отображаемые на графе контролируемых зон определяются двумя SOM-картами. При выборе элемента на SOM-карте «Сотрудники» на графе цветом выделяются вершины, обозначающие зоны, посе-

щенные выбранной группой сотрудников. SOM-карта «Дни» отфильтровывает данные и отображает пространственный шаблон перемещений для заданной группы сотрудников в течение выбранного множества дней.

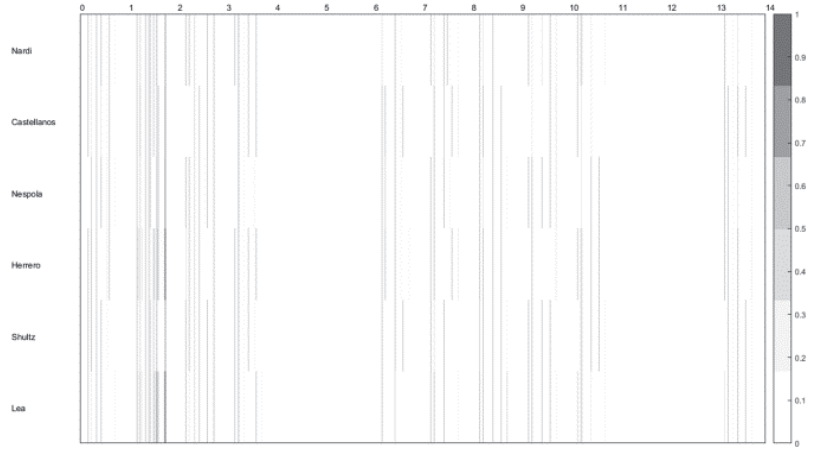
8. Тепловая карта аномалий. Цель тепловой карты — отразить наличие потенциально аномальных отклонений в движении персонала. Аномалии обычно проявляются в нерегулярных нечастых изменениях в поведении объектов, таким образом, предлагается искать отклонения в рамках группы сотрудников или группы дней, имеющих схожие шаблоны перемещений. Тепловая карта строится следующим образом: ось Y соответствует сотрудникам одного кластера, ось X представляет атрибуты векторов, сформированные на этапе подготовки исходных данных. Каждый элемент тепловой карты представляет собой расстояние от значения атрибута вектора признаков сотрудника до центроида соответствующего ему кластера.

Однако непосредственное отображение расстояний может привести к возникновению зашумленной картинки в случае, если расстояния от центроида до объектов внутри кластера примерно одинаковы, а также к исчезновению отклонений на карте, если разница между отдельными расстояниями слишком велика. Например, если сотрудник находится приблизительно 2 часа в определенной зоне в течение заданного интервала времени, тогда отклонение в 10 минут может считаться незначительным, однако в ситуации, когда сотрудник проводит на 10 минут больше в зоне, в которой обычно он находится 20 минут, может иметь место аномалия. Кроме того, оценка отклонений может существенно зависеть от специфики конкретной зоны. Например, отклонение в 20 минут в офисе совершенно несопоставимо с ситуацией, когда сотрудник проводит 20 минут на лестничной площадке.

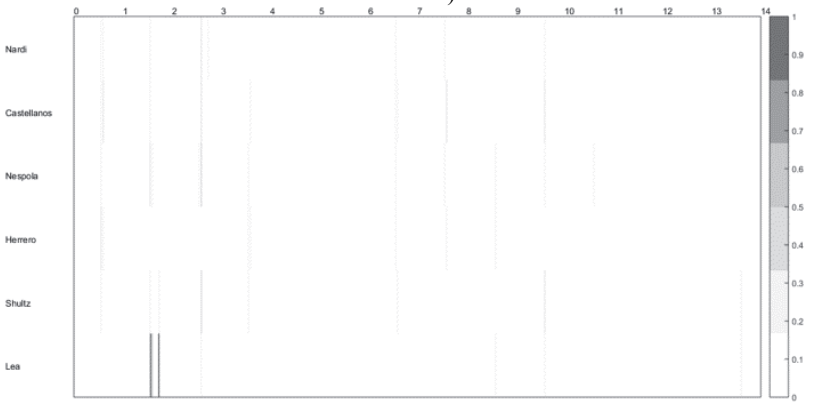
Для решения таких проблем предлагается механизм рейтинговой оценки отклонений, рассматривающий их в контексте некоторого среднего времени пребывания сотрудника в определенной зоне в течение заданного промежутка времени. Основной целью механизма оценки является выбор всех потенциально аномальных отклонений из общего множества и помещение их в фокус внимания аналитика.

Механизм оценки состоит из двух частей — расчет значимости (рейтинга) отклонения и определение пороговых значений для каждой контролируемой зоны. В основе расчета рейтинга отклонения лежит расчет z-показателя (z-score), отражающего расстояние данного значения от среднего значения по набору отклонений. Значения z-показателя распределены на интервале [-4;4] и свидетельствуют, на какую долю стандартного отклонения текущее значение больше или меньше среднего. Значения z-показателя в диапазоне [-1.65;1.65] представляют ожидаемый результат, то есть в такие моменты времени поведение сотрудников почти не отклоняется от их типичного поведения в рамках должностных обязанностей. Значения, выходящие за границы [-2.58;2.58]

свидетельствуют о том, что в данные моменты времени сотрудники демонстрируют нестандартную модель поведения, которая не является результатом случайного процесса [27]. Все отклонения, попавшие за границы данного интервала, можно считать потенциально-аномальными, и поэтому рекомендованы для дальнейшего более подробного изучения. На рисунке 6 продемонстрирован результат применения механизма оценки значимости отклонений в маршрутах сотрудников одного отдела: на рисунке 6а представлены исходные расстояния между значениями атрибутов центроида и объектов данного кластера, на рисунке 6б — расстояния, преобразованные в z-показатель.



а)



б)

Рис. 6. Результат применения механизмов оценки отклонения: а) тепловая карта расстояний между значениями атрибутов центроида и объектов данного кластера; б) тепловая карта z-показателей отклонений

Очевидно, что уровень шума на тепловой карте значительно снизился, и отчетливо стало видно аномальное отклонение в поведении одного из сотрудников группы на второй изучаемый день. Пороговые значения для каждой зоны непосредственно определяют, какие отклонения будут отображаться на тепловой карте, а какие — нет. Фильтрация отклонений производится по их временным и пространственным атрибутам, таким как продолжительность, зона в которой произошло отклонение и соответствующий временной интервал. Таким образом, аналитик может поместить в фокус внимания только отклонения, происходящие на интервале с 8 до 12 часов утра или в холле первого этажа или совместить оба атрибута.

9. Эксперимент и обсуждение результатов. Для оценки предложенного подхода был использован набор данных, представленный в рамках конкурса TheVASTChallenge 2016: Mini-Challenge 2 [10]. Он содержит журнал логов датчиков, регистрирующих появление пользователя в контролируемых зонах здания организации. Когда сотрудник при помощи электронной карты доступа попадает в новую контролируемую зону, данные о его карте автоматически распознаются датчиком и записываются в журнал. Следует отметить, что большая часть зон доступна для сотрудников, даже если они забыли свои пропуска. Журнал содержит записи датчиков за две недели. Аналитик имеет в своем распоряжении схему планировки здания и расположения офисов, включая карты подконтрольных зон, а также список сотрудников с указанием их должностей и офисов.

Исследование траекторий сотрудников организации показало, что большинство сотрудников одного отдела передвигается одинаково, а имеющиеся отличия в моделях поведения можно объяснить спецификой выполняемых задач и расположения рабочего места.

В качестве примера рассмотрим перемещения сотрудников отдела охраны. На рисунке 2 представлена SOM-карта, отражающая результат кластеризации их маршрутов. Типичные маршруты сотрудников отдела представлены на рисунке 7.

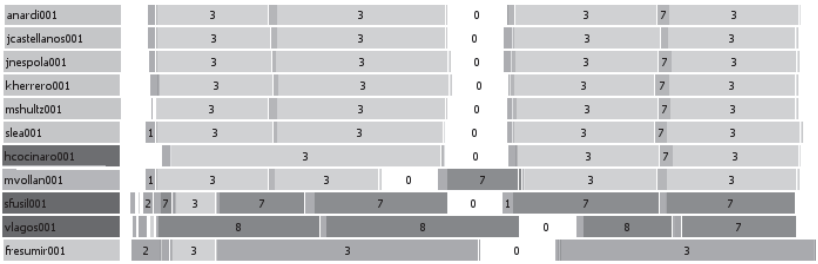


Рис. 7. Типичные маршруты сотрудников отдела «Охрана»

Анализ траекторий членов наиболее многочисленной группы №1, показал, что они имеют два основных шаблона перемещений: один для рабочих дней, другой — для выходных. Их рабочий день обычно начинается около 8:00 и заканчивается в 17:00. Примерно в 12:00 они выходят из здания на час. Большую часть рабочего времени участники этой группы проводят в зоне 2-3 второго этажа, где находятся их рабочие места. Каждые полтора-два часа они покидают свой офис и находятся в смежной зоне 2-7 в течение 5-10 минут. Исходя из того, что охранники посещают только одну зону, расположенную на втором этаже, а также тот факт, что в этой зоне помимо офисов находятся кафе и туалетные комнаты, можно предположить, что они выходят освежиться, не совершая при этом обход здания. В выходные дни они не приходят на работу.

Исследование аномалий в их поведении с помощью тепловой карты позволило обнаружить значительные отклонения в поведении одного из сотрудников на второй день (рисунок 6б). Изучение логов датчиков контроля доступа с помощью модели визуализации Bandview показал отсутствие одного из сотрудников на рабочем месте. Авторы обнаружили еще одну интересную особенность: другой работник группы имеет дублирующиеся записи в журнале для каждого датчика в этот же день, причем вторая временная отметка появляется несколькими секундами позднее первой. Аномалия также хорошо заметна на графе посещаемых зон, отображающего статистику посещений за день для каждой из зон. Факт отсутствия записей в журнале для одного сотрудника и дублирование для другого позволяет предположить, что первый сотрудник использовал электронный пропуск второго сотрудника для прохода в контролируемую зону.

Два ближайших к этой группе кластера (группы №2 и №3) состоят из одного человека. Их маршруты достаточно похожи на маршруты сотрудников группы №1, поскольку их офисы также расположены в зоне 2-3, где они и проводят большую часть времени. Основное отличие для сотрудника группы №2 заключается в том, что он не покидает зону 2-3 в первой половине дня и посещает зону 2-7 только во второй половине дня около 15:00. Исследование его перемещений с помощью модели визуализации BandView и тепловой карты позволило выявить небольшое отклонение, произошедшее на третий день всего контролируемого периода. В этот день во время послеобеденного обхода зоны 2-7 он не возвращается в свою комнату, а в 17:00 идет домой из зоны 2-7. Это отклонение можно объяснить тем, что сотрудник мог забыть приложить электронный пропуск, возвращаясь на рабочее место из зоны 2-7. Цвет ячейки SOM-карты, соответствующей кластеру №3, значительно темнее, и различия в

траекториях выражены сильнее. Данный сотрудник отправляется обедать на час раньше, а с 12:00 до 13:00 находится в зоне 1-7. Каких-либо значимых отклонений у данного сотрудника выявлено не было.

Оставшиеся две группы (№4 и №5) расположены отдельно от остальных. Группа №4 состоит из двух сотрудников, чьи офисы располагаются на первом этаже. Они имеют 3 шаблона маршрутов, зависящих от дня недели, представленные на рисунке 8. Во все рабочие дни, кроме каждого вторника, они проводят большую часть времени на рабочем месте, расположенном в зоне 1-7. Каждый полтора-два часа совершают обход зоны 1-1, а в обеденное время с 12:00 до 13:00 покидают здание. Каждый вторник в первой половине дня они посещают зону 2-3, где располагаются офисы их коллег, что позволяет предположить, что они посещают еженедельные собрания, на которых проходят очередной инструктаж. Авторы обнаружили два интересных отклонения в перемещениях одного из сотрудников этого кластера. Данный сотрудник посетил зону 3-4, расположенную на третьем этаже, только один раз за весь наблюдаемый двухнедельный промежуток времени, причем длительность пребывания в этой зоне не превысила 1 минуты. Вторая аномалия связана с нетипичной длительностью пребывания в зоне 1-2. Проводя обычно в этой зоне около 2-3 минут, он в один из дней провел в ней 15 минут. Поскольку в этой зоне находится кафе, можно предположить, что обычно он совершает обход помещений первого этажа, в то время как во время длительного пребывания в зоне 2-3 он общался с кем-то из сотрудников, также находившихся в этой зоне. Однако для уточнения этого предположения следует изучить, кто еще в это время находился в зоне 2-3.

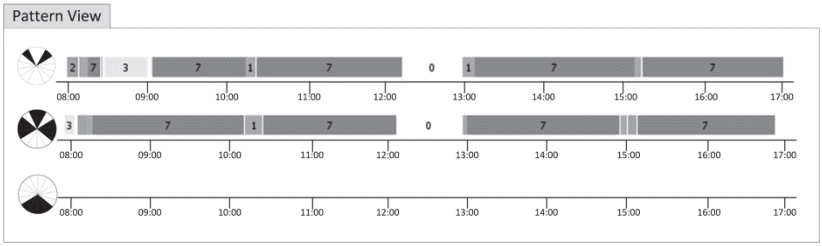


Рис. 8. Шаблоны маршрутов сотрудников группы №4

Группа №5 состоит из одного сотрудника. Большую часть времени он проводит на третьем этаже, где находится его офис. Тот факт, что на третьем этаже располагаются офисы руководства компании, позволяет предположить, что этот сотрудник является начальником отдела безопасности. Его перемещения достаточно разнообразны и сильно зависят от дня недели. Он начинает свой рабочий день с посе-

щения местного кафе, находящегося в зоне 1-2, и затем поднимается в свой офис. Каждый вторник он заходит в зону 2-3 и проводит там около получаса, а затем возвращается в свой офис. Анализ результатов представления исходных данных о перемещениях сотрудников отдела безопасности, полученный с помощью модели визуализации BandView, позволяет заключить, что каждые вторник руководитель отдела проводит совещания со всем отделом. Кроме того, можно установить, что представитель отдела безопасности должен постоянно находиться на первом этаже с восьми часов утра до пяти часов вечера. По этой причине некоторые из сотрудников отдела выходят на обеденный перерыв в другое время; а работники, чьи офисы располагаются на первом этаже, посещают собрания только раз в неделю, чтобы обеспечить присутствие охраны на входе в здание.

Предложенный подход позволил описать основные модели поведения сотрудников организации, определить существующие ограничения в их перемещениях. Использование тепловой карты в сочетание с оценкой отклонений в маршрутах позволяет достаточно легко обнаружить аномалии разного рода, начиная от отсутствия сотрудника на рабочем месте и заканчивая дублированием логов и нетипичной длительностью нахождения в контролируемой зоне.

Следует отметить, что в большинстве случаев анализ контактов персонала способствует лучшему пониманию причин возникновения аномалий. Модель BandView может быть полезна для понимания возможного взаимодействия между работниками. Однако она работает лишь в том случае, когда число отображаемых работников ограничено и не превышает 10-15 человек, иначе отслеживание взаимодействий между сотрудниками сильно затруднено.

Таким образом, одно из основных направлений будущей работы связано с разработкой методики визуального анализа, позволяющей установить паттерны взаимодействия между сотрудниками организации. Возможными способами решения данной проблемы является построение графов контактов, отражающие структурные особенности взаимодействия сотрудников, тепловые карты посещения зон различными группами сотрудников для выявления различных типов взаимодействия. Другое направление будущей работы затрагивает анализ данных, полученных из различных источников. Журналы операционной системы, например события входа-выхода и события клавиатуры, доказывают, что сотрудник находится на рабочем месте. Показания сенсоров инфраструктур здания, например, таких как тепло-вентиляционная система, могут также объяснить нетипичное поведение сотрудников. Корреляция этих данных требует разработки

новых методов визуального анализа с учетом особенностей исходных данных.

10. Заключение. В статье представлен подход к анализу траекторий сотрудников критических инфраструктур, основанный на использовании методик визуализации и автоматического анализа данных. Он позволяет сформировать пространственно-временные шаблоны маршрутов, установить имеющиеся ограничения и обнаружить аномалии в перемещениях сотрудников. Ключевыми элементами подхода являются интерактивные SOM-карты, которые используются для выявления групп сотрудников, имеющих схожие траектории перемещений, и установления возможной периодичности в их маршрутах. Выявление аномалий в перемещениях сотрудников осуществляется с помощью тепловой карты, дополненной механизмом оценки отклонений в контексте пространственно-временного шаблона поведения сотрудника. Авторы представили методики взаимодействия, связывающие все визуальные компоненты и обеспечивающие процесс анализа. Для иллюстрации предложенного подхода был использован набор данных, предоставленный в рамках конкурса VastChallenge 2016. В статье представлено обсуждение полученных результатов и определены основные направления будущей работы.

Учитывая требования к исходным данным — наличие метки времени посещения контролируемой зоны, идентификатор сотрудника, идентификатор контролируемой зоны, данная методика естественным образом может быть добавлена в систему управления и контроля доступом в качестве модуля, выполняющего интеллектуальный анализ данных. В связи с ростом актуальности задачи выявления внутреннего нарушителя она может быть реализована в компоненте визуального анализа поведения сотрудников организации систем управления информационной безопасностью при условии включения датчиков системы доступа в качестве сенсоров данных, обрабатываемой SIEM-системой [28, 29].

Литература

1. *Demšar U. et al.* Analysis and visualisation of movement: an interdisciplinary review // Movement Ecology. 2015. vol. 3. no. 1.
2. *Lerman Y., Rofe Y., Omer I.* Using Space Syntax to Model Pedestrian Movement in Urban Transportation Planning // Geographical Analysis. 2014. vol. 46(4). pp. 392–410.
3. *Ferreira N. et al.* Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips // IEEE Transactions on Visualization and Computer Graphics. 2013. vol. 19. pp. 2149–2158.
4. *Tan L., Hu M., Lin H.* Agent-based simulation of building evacuation // International Journal of Information Sciences. 2015. vol. 295. no. C. pp. 53–66.
5. Hubstuff Employee Monitoring Software. URL: https://hubstaff.com/employee_monitoring_software (дата обращения: 29.07.2017).

6. WaveTrend Access Control. URL: http://www.wavetrend.net/access-control.php (дата обращения: 29.07.2017).
7. Employee Monitoring and Productivity Analysis. URL: https://www.intesecurity.com/employee-monitoring-and-productivity-analysis/ (дата обращения: 29.07.2017).
8. ObserveIT Insider Threat Solution. URL: https://www.observeit.com/insider-threat-solution (дата обращения: 29.07.2017).
9. *Bussa T., Litan A., Phillips T.* Market Guide for User and Entity Behavior Analytics. URL: https://www.gartner.com/doc/3538217/market-guide-user-entity-behavior (дата обращения: 29.07.2017).
10. Vast Challenge Website. URL: http://vacommunity.org/ (дата обращения: 05.04. 2017).
11. *Kisilevich S., Mansmann F., Nanni M., Rinzivillo S.* Spatio-temporal clustering: a survey // Data Mining and Knowledge Discovery Handbook. 2010. pp. 855–874.
12. *Andrienko G., Andrienko N.* Exploration of massive movement data: a visual analytics approach // Proceedings of the 11th AGILE International Conference on Geographic Information Science. 2008. URL: https://agile-online.org/conference_paper/cds/agile_2008/pdf/66_doc.pdf (дата обращения: 29.07.2017).
13. *Kotenko I., Novikova E.* VisSecAnalyzer: a Visual Analytics Tool For Network Security Assessment // International Conference on Availability, Reliability, and Security. 2013. LNCS 8128. pp. 345–360.
14. *Schreck T., Bernard J., Von Landesberger T., Kohlhammer J.* Visual cluster analysis of trajectory data with interactive Kohonen map // Information Visualization. 2009. vol. 8. no. 1. pp.14–29.
15. *Guo D., Chen J., MacEachren A.M., Liao K.* A visualization system for space-time and multivariate patterns (VIS-STAMP) // IEEE Transactions on Visualization and Computer Graphics. 2006. vol. 12(6). pp. 1461–1474.
16. *Andrienko N., Andrienko G.* Visual analytics of movement: an overview of methods, tools and procedures // Information Visualization. 2013. vol. 12(1). pp. 3–24.
17. *Andrienko G. et al.* Scalable Analysis of Movement Data for Extracting and Exploring Significant Places // IEEE Transactions on Visualization and Computer Graphics. 2013. vol. 19. pp. 1078–1094.
18. *Ho Q., Nguyen P.H, Åström T., Jern M.* Implementation of a Flow Map Demonstrator for Analyzing Commuting and Migration Flow Statistics Data // Procedia – Social and Behavioral Sciences. 2011. vol. 21. pp. 157–166.
19. *Abel J., Sander N.* Quantifying Global International Migration Flows // Science. 2014. vol. 343. Issue 6178. pp. 1520–1522.
20. *Andrienko G., Andrienko N., Schumann H., Tominski C.* Visualization of Trajectory Attributes in Space–Time Cube and Trajectory Wall // Cartography from Pole to Pole. 2014. pp. 157–163.
21. *Guo C. et al.* Dodeca-Rings Map: Interactively Finding Patterns and Events in Large Geo-temporal Data // Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST). 2014. pp. 353–354.
22. *Choo J. et al.* Exploring Anomalies in GAStech: VAST Mini Challenge 1 and 2 // Proceedings of IEEE Symposium on Visual Analytics Science and Technology Challenge (VAST). 2014. pp. 347–348.
23. *Tominski C., Schumann H., Andrienko G., Andrienko N.* Stacking-Based Visualization of Trajectory Attribute Data // IEEE Transactions on Visualization and Computer Graphics. 2012. vol. 18. no. 12. pp. 2565–2574.
24. *Shneiderman B.* Dynamic queries for visual information seeking // IEEE Software. 2003. vol. 11. no. 6. pp.70–77.

25. *Kohonen T., Honkela T.* Kohonen network // Scholarpedia. 2007. vol. 2(1). pp. 1568.
26. *Ultsch A.* Self-organizing neural networks for visualization and classification // Information and Classification. pp. 307–313.
27. *Caldas de Castro M., Singer B.* Controlling the False Discovery Rate: A New Application to Account for Multiple and Dependent Test in Local Statistics of Spatial Association // Geographical Analysis. 2006. vol. 38. pp. 180–208.
28. *Котенко И.В., Саенко И.Б.* Построение системы интеллектуальных сервисов для защиты информации в условиях кибернетического противоборства // Труды СПИИРАН. 2012. Вып. 3(22). С. 84–100.
29. *Kotenko I., Chechulin A.* Attack Modeling and Security Evaluation in SIEM Systems // International Transactions on Systems Science and Applications. 2012. vol. 8. pp.129–147.

Новикова Евгения Сергеевна — к-т техн. наук, старший научный сотрудник лаборатории проблем компьютерной безопасности Федеральное государственное бюджетное учреждение науки Санкт-Петербургского института информатики и автоматизации Российской академии наук (СПИИРАН), доцент кафедры информационных систем, Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (СПбГЭТУ). Область научных интересов: визуальная аналитика, вредоносное программное обеспечение, двухключевая криптография. Число научных публикаций — 80. novikova.evgenia123@gmail.com, http://www.comsec.spb.ru/en/staff/novikova; 14-я линия В.О., 39, Санкт-Петербург, 199178; р.т.: +7(812)328–2642.

Муренин Иван Николаевич — студент магистратуры, Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (СПбГЭТУ). Область научных интересов: анализ траекторий, визуализация многомерных данных. Число научных публикаций — 2. imurenin@gmail.com; ул. Профессора Попова, 5, Санкт-Петербург, 197376; р.т.: +7(812) 234-27-73.

Поддержка исследований. Работа выполнена при финансовой поддержке РФФИ (проект № 16-07-00625) в СПБГЭТУ и при частичной поддержке бюджетных тем № 0073-2015-0004 и 0073-2015-0007 в СПИИРАН.

.

DOI 10.15622/sp.54.3

E.S. NOVIKOVA, I.N. MURENIN

THE TECHNIQUE OF THE VISUAL ANALYSIS OF THE ORGANIZATION EMPLOYEES ROUTES FOR ANOMALY DETECTION

Novikova E. S., Murenin I. N. **The Technique of the Visual Analysis of the Organization Employees Routes for Anomaly Detection.**

Abstract. The detection of anomalies in the movement of employees is an important task of the cyber-physical security of enterprises, including critical infrastructures. The paper presents a technique to analyze the routes of the organization employees based on combination of the data mining and interactive visualization techniques. It includes two stages – detection of the groups of the employees with similar behavior and anomaly discovery. The self-organizing Kohonen maps are used to group employees on the basis of their behavior. To present spatio-temporal patterns, authors developed special visualization model named BandView. To detect anomalies authors present a rating mechanism assessing spatiotemporal attributes of the movement. The visualization of the anomalies is done using heatmaps that allow an analyst to spot place and time with a possibly suspicious activity. The technique is tested against data set provided within VAST MiniChallenge-2 contest that contains logs from access control sensors describing employees' movement within organization building.

Keywords: anomaly detection in trajectories, visual analytics, behavior patterns, behavior deviation assessment, heatmaps.

Novikova Evgenia Sergeevna — Ph.D., senior researcher of the computer security problems laboratory, St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences (SPIIRAS), associate professor of the information systems department , Saint-Petersburg State Electrotechnical University "LETI" (ETU). Research interests: security visual analytics, malware, public key cryptography. The number of publications — 80. evgeshka19@mail.ru, http://www.comsec.spb.ru/en/staff/novikova; 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone: +7(812)328–2642.

Murenin Ivan Nikolaevich — trajectories analysis, multidimensional data visualization. Research interests: master student Saint-Petersburg State Electrotechnical University "LETI" (ETU). The number of publications — 2. imurenin@gmail.com; 5, Professor Popov str., St.Petersburg, 197376, Russia; office phone: +7(812) 234-27-73.

Acknowledgements. This research is supported by RFBR (projects No. 16-07-00625 in ETU and partly by the budget (projects no. 0073-2015-0004 and 0073-2015-0007) in SPIIRAS.

References

1. Demšar U., Buchin K., Cagnacci F., Safi K., Speckmann B., Van de Weghe N., Weibel R. Analysis and visualisation of movement: an interdisciplinary review. *Movement Ecology*. 2015. vol. 3. no. 1.
2. Lerman Y., Rofe Y., Omer I. Using Space Syntax to Model Pedestrian Movement in Urban Transportation Planning. *Geographical Analysis*. 2014. vol. 46(4). pp. 392–410.
3. Ferreira N. et al. Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips. *IEEE Transactions on Visualization and Computer Graphics*. 2013. vol. 19. pp. 2149–2158.

4. Tan L., Hu M., Lin H. Agent-based simulation of building evacuation. *International Journal of Information Sciences*. 2015. vol. 295. no. C. pp. 53–66.
5. Hubstuff Employee Monitoring Software. Available at: https:.hubstaff.com/employee_monitoring_software (accessed: 29.07.2017).
6. WaveTrend Access Control, URL: http:.www.wavetrend.net/access-control.php (accessed: 29.07.2017).
7. Employee Monitoring and Productivity Analysis. Available at: https:.www.intesecurity.com/employee-monitoring-and-productivity-analysis/ (accessed: 29.07.2017).
8. ObserveIT Insider Threat Solution. Available at: https:.www.observeit.com/insider-threat-solution (accessed: 29.07.2017).
9. Bussa T., Litan A., Phillips T. Market Guide for User and Entity Behavior Analytics. Available at: https:.www.gartner.com/doc/3538217/market-guide-user-entity-behavior (accessed: 29.07.2017).
10. Vast Challenge Website. Available at: http:.vacommunity.org/ (accessed: 05.04. 2017).
11. Kisilevich S., Mansmann F., Nanni M., Rinzivillo S. Spatio-temporal clustering: a survey. Data Mining and Knowledge Discovery Handbook. 2010. pp. 855–874.
12. Andrienko G., Andrienko N. Exploration of massive movement data: a visual analytics approach. Proceedings of the 11th AGILE International Conference on Geographic Information Science. 2008. Available at: https://agile-online.org/conference_paper/cds/agile_2008/pdf/66_doc.pdf (accessed: 29.07.2017).
13. Kotenko I., Novikova E. VisSecAnalyzer: a Visual Analytics Tool For Network Security Assessment. International Conference on Availability, Reliability, and Security. 2013. LNCS 8128. pp. 345–360.
14. Schreck T., Bernard J., Von Landesberger T., Kohlhammer J. Visual cluster analysis of trajectory data with interactive Kohonen map. *Information Visualization.* 2009. vol. 8. no. 1. pp.14–29.
15. Guo D., Chen J., MacEachren A.M., Liao K. A visualization system for space-time and multivariate patterns (VIS-STAMP). *IEEE Transactions on Visualization and Computer Graphics.* 2006. vol. 12(6). pp. 1461–1474.
16. Andrienko N., Andrienko G. Visual analytics of movement: an overview of methods, tools and procedures. *Information Visualization.* 2013. vol. 12(1). pp. 3–24.
17. Andrienko G. et al. Scalable Analysis of Movement Data for Extracting and Exploring Significant Places. *IEEE Transactions on Visualization and Computer Graphics.* 2013. vol. 19. pp. 1078–1094.
18. Ho Q., Nguyen P.H, Åström T., Jern M. Implementation of a Flow Map Demonstrator for Analyzing Commuting and Migration Flow Statistics Data. *Procedia - Social and Behavioral Sciences*. 2011. vol. 21. pp. 157–166.
19. Abel J., Sander N. Quantifying Global International Migration Flows. *Science.* 2014. vol. 343. Issue 6178. pp. 1520–1522.
20. Andrienko G., Andrienko N., Schumann H., Tominski C. Visualization of Trajectory Attributes in Space–Time Cube and Trajectory Wall. Cartography from Pole to Pole. 2014. pp. 157–163.
21. Guo C. et al. Dodeca-Rings Map: Interactively Finding Patterns and Events in Large Geo-temporal Data. Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST). 2014. pp. 353–354.
22. Choo J. et al. Exploring Anomalies in GAStech: VAST Mini Challenge 1 and 2. Proceedings of IEEE Symposium on Visual Analytics Science and Technology Challenge (VAST). 2014. pp. 347–348.
23. Tominski C., Schumann H., Andrienko G., Andrienko N. Stacking-Based Visualization of Trajectory Attribute Data. *IEEE Transactions on Visualization and Computer Graphics.* 2012. vol. 18. no. 12. pp. 2565–2574.

24. Shneiderman B. Dynamic queries for visual information seeking. *IEEE Software.* 2003. vol. 11. no. 6. pp.70–77.
25. Kohonen T., Honkela T. Kohonen network . *Scholarpedia.* 2007. vol. 2(1). pp. 1568.
26. Ultsch A. Self-organizing neural networks for visualization and classification. *Information and Classification*. pp. 307–313.
27. Caldas de Castro M., Singer B. Controlling the False Discovery Rate: A New Application to Account for Multiple and Dependent Test in Local Statistics of Spatial Association. *Geographical Analysis*. 2006. vol. 38. pp. 180–208.
28. Kotenko I.V., Saenko I.B. [Developing the system of intelligent services to protect information in cyber warfare]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2012. vol. 3(22). pp. 84–100. (In Russ.).
29. Kotenko I., Chechulin A. Attack Modeling and Security Evaluation in SIEM Systems. *International Transactions on Systems Science and Applications.* 2012. vol. 8. pp. 129–147.