УДК 004.056.5

П.И. Банокин

Модель поведения пользователя корпоративной информационной системы

Представлена модель поведения пользователя корпоративной информационной системы на основе графа де Брюина второго порядка и ассоциированных с вершинами графа признаков текстовых данных. Текстовые признаки созданы с учетом специфики данных, отображаемых в клиентских приложениях информационных систем. Модель позволяет определить отклонения в поведении, которые в том числе могут быть вызваны внутренними утечками данных. Определяемые с помощью модели отклонения включают сценарии выгрузки данных несвойственной тематики и доступа к данным типичной тематики, но с нарушением исполнения бизнес-процессов. Модель оценена на основе критериев полноты, адекватности и экономичности.

Ключевые слова: внутренние утечки данных, корпоративные информационные системы, модель поведения.
DOI: 10.21293/1818-0442-2023-26-4-78-83

Внутренняя утечка данных – это утрата конфиденциальности данных, произошедшая в результате действий сотрудника предприятия [1]. Рост активности преступников в области информационной технологий [2] и риски финансовых убытков [3] обусловливают актуальность защиты корпоративных данных. Согласно публикациям, количество случаев утечек корпоративных данных имеет устойчивую ежегодную тенденцию к увеличению [4]. Утечки данных наносят репутационный и экономический вред. По способу осуществления утечки данных разделяются на две категории: внешние и внутренние. В реализации внутренних утечек данных участвует сотрудник компании, который может выполнять заказ преступника на поставку данных [5]. Сотрудник имеет авторизованный доступ к интересующим заказчика данным и может быть хорошо осведомлен о мероприятиях и средствах обеспечения информационной безопасности. Одним из способов противодействия внутренним утечкам данных является использование специализированного программного обеспечения для анализа поведения пользователей. Модули анализа поведения пользователей предоставляются в различных пакетах DLP [6], но их исследование невозможно из-за закрытого исходного кода и лицензионных соглашений.

Поведение – это устоявшаяся система действий пользователя, учитывающая время, текущее и предыдущие состояния клиента КИС и возможные внешние воздействия. Модель поведения включает описание системы действий пользователя, позволяющее с помощью алгоритма идентификации утечек данных оценить степень нормальности (соответствия модели) действий пользователя.

Одной из ранних работ, посвященных анализу поведения и противодействию внутренним утечкам данных, является публикация [7]. В работе в качестве источника данных рассматривается коллекция лог-записей, используемая для формирования профилей поведения на основе статистических моделей. Формат лог-записи включает данные о пользователе, команде и объектах. В дальнейшем с увеличением сложности программного обеспечения и количества типов действий и событий для анализа поведения стали применяться методы обработки естественного языка и методы теории графов.

Методы обработки естественного языка включают метрики обратной частоты документа (inversed document frequency, IDF) [8, 9], применяемые для категориальных данных – типов событий, и векторные представления текстовых данных [10]. В работе [8] выявляются необычные частоты событий и необычное время их возникновения для обнаружения аномалий. В процессе выявления аномалий используется фиксированный перечень ранжированных по важности событий с применением метрики обратной частоты документа (IDF), в роли текстового документа выступает рабочий день сотрудника предприятия, содержащий перечень событий и сопоставленные им время и количественные значения. В исследовании [11] анализ текстовых данных лог-записей ограничен описанием события и базой синонимов и антонимов событий.

При моделировании поведения пользователей графовые модели создаются как для одного пользователя, так и для множества пользователей и программных ресурсов, в том числе с использованием двудольного графа [12]. Узлы графа с наибольшей степенью и ребра графа с максимальным весом присутствуют в последовательностях лог-записей, относящихся к классу нормального поведения [13]. В работе [13] степень нормальности поведения находится по категории подграфа, которая определяется рангом вершин и весом ребер. В случае анализа действий пользователя такая оценка накладывает ограничение на длину анализируемой подпоследовательности (окна) действий, так как с увеличением длины возрастает вероятность наличия шумовых данных из-за случайных ошибок в действиях сотрудника. Единичный случай посещения редкой вершины является маловероятным признаком осуществленной утечки данных, но посещение безопасных вершин безопасного подграфа может не учитывать нарушения бизнес-процессов. В задаче идентификации утечек данных более

предпочтительно создание моделей поведения для каждого пользователя отдельно, что позволит учитывать индивидуальные особенности поведения сотрудника с отбором информативных признаков. В работе [14] проводится совместный анализ последовательностей действий и вычисленных атрибутов, но не учитываются семантические признаки текстовых данных.

По способу реализации контроля за действиями пользователя можно выделить два подхода: наблюдение на сервере-источнике данных и наблюдение на конечном устройстве пользователя. При первом подходе упрощается архитектура системы предотвращения утечек данных, но возможны ограничения, накладываемые на процесс разработки программы-прокси для отслеживания сообщений – вызовов методов интерфейса программирования, из-за закрытого исходного кода КИС и системы управления базами данных. При втором подходе необходима установка программы-агента для наблюдения за действиями пользователя на клиентские устройства, но получаемые лог-записи могут включать более широкий набор атрибутов, в том числе графовые и текстовые данные. Текстовая составляющая представлена загруженными пользователем для просмотра записями корпоративной БД, а графовая – последовательностями перехода между элементами графического интерфейса клиента КИС. В статье в дальнейшем будет рассматриваться модель поведения при наблюдении на клиентских устройствах. Анализ данных графовой модели позволит определять нарушения в исполнении бизнес-процессов, а анализ текстовых данных – выявить обращения к данным несвойственной тематики.

В упомянутых выше работах не выполняется совместный анализ данных графовой и документной моделей, а также не учитывается специфика работы пользователя с КИС. Важными требованиями к модели являются объясняющая способность и возможность поиска поведенческих аномалий в коротких последовательностях действий. Объясняющая функция выражена в возможности получить характеристики поведения пользователя, включая перечень действий и признаки текстовых документов в момент обнаружения поведенческой аномалии. Поиск аномалий с использованием коротких последовательностей (окон) необходим для предотвращения хищений небольших выборок данных.

Целью работы является создание модели поведения пользователя КИС, удовлетворяющей представленными выше требованиям, и ее оценка критериями универсальности, адекватности и экономичности. В рамках исследования использованы методы теории графов и машинного обучения.

Новизна настоящей работы заключается в использовании графа де Брюина для создания модели поведения, хранящей ассоциированные со сменой состояний программы коллекции текстовых документов и их признаков, и алгоритма расчета веса вершин графа, учитывающего специфику работы с КИС. Особенностью программной реализации является получение текстовых данных корпоративных информационных систем на стороне клиента, что позволяет создавать лог-записи для облачных ERP-систем.

Модель поведения пользователя

При совершении пользователем действия при работе с интерфейсом клиента КИС создается лог-запись x_i = (user, timestamp, text, uiElement, url), которая является вектором с компонентами, содержащими данные об идентификаторе пользователя, времени события, текстовых данных клиента КИС, идентификаторе элемента интерфейса и URL-адресе. Дано множество пользователей КИС $\mathbf{U} = \{u_i\}$. Пусть дана для каждого пользователя упорядоченная по времени поступления последовательность лог-записей $P_{u_i} = (x_1, x_2, ..., x_n)$ и задано множество состояний клиента КИС $\mathbf{S} = \{s_i\}$. Коллекция P_{u_i} содержит данные о нормальном поведении пользователя: посторонние элементы отсутствуют или содержатся в крайне незначительном количестве. Состояние s_i клиента КИС идентифицируется элементом управления интерфейса пользователя, при воздействии на который происходит изменение текстового содержимого клиента КИС. В случае веб-приложений таким идентификатором является значение xpath HTML-элемента интерфейса или часть этого значения при объединении нескольких элементов в один узел.

При наблюдении на клиентском устройстве пользователя значение атрибута text лог-записи x_i включает все параграфы текста, отображаемые в графическом клиенте КИС. Для процесса анализа поведения необходимы только новые данные, полученные из корпоративной БД при смене состояния программы в результате совершения пользователем действия. Новые текстовые данные определяются как разность множеств

$$v_i^{<\text{pageText}>} = \mathbf{P}_{k+1} / \mathbf{P}_k \neq \varnothing,$$

где $\mathbf{P}_k$ – множество параграфов текста в состоянии программы в момент k, $\mathbf{P}_{k+1}$ – множество параграфов текста в состоянии программы в некоторый момент времени k+1, v_i – вершина поведенческого графа.

Вычисление разности текстовых данных позволит получить только те параграфы текста, которые были найдены в результате изменения состояния прикладной программы. Кроме этого, из анализа будут исключены текстовые значения повторяющихся элементов интерфейса, включая меню, рубрикатор, строки состояния и другие элементы. При представлении перехода программы из состояния s_k в состояние s_{k+1} в виде объединённого узла $v(s_k s_{k+1})$ возможно построение графа де Брюина [15] второго порядка (рис. 2). Преимуществами использования графа второго порядка являются не только удобство получения измененных текстовых данных, но и хранение информации о предыдущем состоянии программы (контекста) в идентификаторе вершины. Использование графов более высокого порядка значительно увеличивает количество признаков-вершин и вычислительную сложность.

Таким образом, на основе накопленных лог-записей строится модель поведения, представленная графом второго порядка с вершинами, атрибуты которых хранят данные документной модели – текстовые данные и их признаки (рис. 1). Для создания поведенческого графа второго порядка дана коллекция лог-записей P_{u_i}, каждому элементу которой сопоставляется состояние из множества **S**. Узлы взвешенного поведенческого графа $\mathbf{G}=(V,E)$ представляют собой не отдельные состояния, а два связанных состояния $v_i=(s_t,s_{t+1})$. Одновременно с построением графа выполняется проход коллекции лог-записей P_{u_i} окном window длины l из множества перекрывающих окон Windows. Значение параметра l соответствует среднему количеству действий пользователя, необходимого для исполнения единицы бизнес-процесса. В работе используется параметр l=24. Каждой вершине графа соответствуют два счетчика посещений: счетчик присутствия в окнах count_window и абсолютный счетчик присутствия в последовательности P_{u_i} count_total.

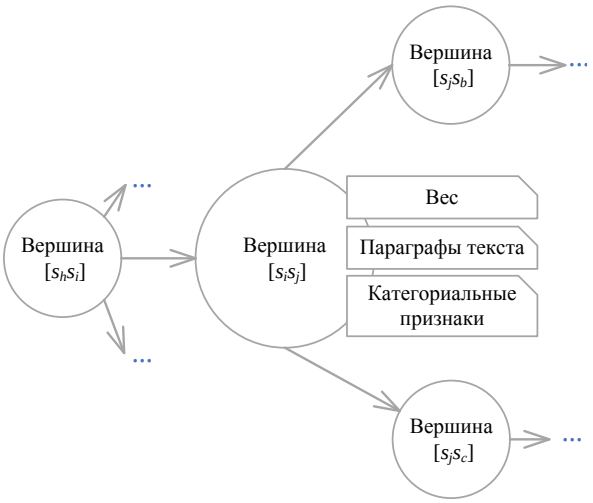

Рис. 1. Схема поведенческого графа

Значение count_window рассчитывается как сумма количества посещений вершины в каждом из временных окон:

$$v_i^{<\text{count\_window}>}=\sum_{|\text{Windows}|}\text{count}(v_i,\text{window}),$$

где count(v_i, windows) – количество посещений вершины v_i в окне window.

Итоговый вес узла v_i рассчитывается по формуле и учитывает значимость узла по присутствию в последовательностях действий, понижая вес повторяющихся действий при исполнении бизнес-процессов:

$$\text{weight}(v_i)=\frac{v_i^{<\text{count\_total}>}}{v_i^{<\text{count\_window}>}}.$$

Также узлы v_i графа **G** имеют атрибуты text_features, pageText, times. Атрибут узла $v_i^{<\text{text\_features}>}$ содержит коллекцию признаков текстовых данных, вычисленных на основе коллекции параграфов $v_i^{<\text{page\_text}>}$. Атрибут узла $v_i^{<\text{times}>}$ содержит коллекцию интервалов времени, в течение которых программа находилась в вершине v_i. Интервалы времени могут быть использованы как косвенный признак изменений поведения при подробном изучении обнаруженной аномалии. Вес ребра графа определяется отношением количества переходов по ребру edge(v_i,v_j) к общему числу переходов в вершину v_j:

$$\text{weight}(v_i,v_j)=\frac{\text{count}(v_i,v_j)}{v_j^{<\text{count\_total}>}},$$

где count(v_i, v_j) – количество переходов из вершины v_i в вершину v_j.

Текстовые данные КИС создаются сотрудниками или поступают в корпоративную БД от сторонних организаций или программных сервисов. Текстовые документы могут быть созданы на основе типовых форм и лишены грамматической структуры и авторского стиля. В ряде случаев параграфы текста могут представлять собой короткие текстовые метки элементов интерфейса. Эти особенности делают невозможным применение методов определения авторства текста, включающих исследование стилистики текста, лексического разнообразия и других авторских особенностей. Поэтому необходимо при создании модели учитывать возможность проверки степени схожести новых текстовых документов с ранее наблюдаемыми текстовыми данными на основе категориальных признаков, полученных из текста. Примерами специфичных для корпоративного использования признаков являются следующие характеристики текста: почтовые индексы, телефонные номера, валютные коды, товарные категории, названия юридических лиц, географические коды и др.

Пусть задано множество категориальных признаков $\mathbf{Z}=\{Z_1,Z_2,...,Z_n\}$, вычисляемых в коллекции параграфов вершин $v_i^{<\text{page\_text}>}$ графа с множеством значений $\mathbf{C}=\mathbf{C}_1\cup\mathbf{C}_2\cup...\cup\mathbf{C}_n$. Каждый признак Z_i имеет соответствующее ему множество значений $\mathbf{C}_i=\{c_0^i,c_2^i,c_2^i,...,c_{|C_i|}^i\}$ и функцию-гистограмму распределения этих значений $\text{hist}(c_j^i,\mathbf{C}_i)\to\mathbf{R}$, полученную на основе выборки P_{u_i}. Значение $c_0{}^i$ выполняет роль индикатора наличия значения, ранее не наблюдаемого в накопленной коллекции P_{u_i}.

Перед оценкой безопасности поведения пользователя происходит создание объекта модели – поведенческого графа второго порядка с вычисленными весами вершин и атрибутами. Оценка безопасности поведения производится с использованием последовательности новых лог-записей (временных окон) window = (x_t, x_{t+1}, ..., x_{t+l}) длины l переходов по вершинам поведенческого графа. Для выбора параметра l необходимо учитывать характер обнаруживаемых аномалий, которые по мере увеличения длины окна могут включать ошибки при работе с интерфейсом программы (несколько переходов по ребрам с низким

весом), кратковременные нарушения бизнес-процессов (выгрузка небольших выборок данных) и последовательную выгрузку данных во время перерывов в работе сотрудника. Окно window преобразуется в вектор окна $w = (\text{weight}(v_1)*v_1, ..., \text{weight}(v_m)*v_m, c_1, ..., c_{|\mathbf{C}|})$, где компоненты $\text{weight}(v_i)*v_i$ содержат количества посещений вершины v_i с учетом ее веса $\text{weight}(v_i)$ в графе **G**, а компоненты c_i – количества обнаруженных значений признаков из множества **C**. Значения компонент вектора w нормализуются отдельно для каждого подмножества значений признака Z_i. Подмножества компонент вектора w обрабатываются функциями поиска аномалий. Отдельная обработка подмножеств позволяет объяснить причину поведенческой аномалии. При обнаружении аномалии по несоответствию тематики происходят детализация на основе атрибутов вершин подграфа $\mathbf{G}_{\text{window}}$ и указания несоответствия значений признаков ранее наблюдаемому распределению признаков вершин. При нарушении типичного исполнения бизнес-процессов происходит детализация с указанием весов ребер, степени вершин и продолжительности посещения вершин.

Оценка соответствия последовательности window поведенческому графу производится с помощью функции поиска аномалий, реализация которой может быть выполнена на основе нейронной сети архитектуры автокодировщик, основе ансамбля классификаторов на основе сравнения гистограмм распределения значений и других алгоритмов поиска посторонних значений.

Оценка модели

Модель оценивается критериями универсальности, адекватности и экономичности. Представленная модель учитывает текстовые данные КИС и последовательности действий пользователя, тем самым является более универсальной по сравнению с моделями, учитывающими только текстовые данные или только последовательности действий и событий.

Экономичность модели обеспечивается хранением только измененных текстовых данных и вычисленных признаков и включением в поведенческий граф второго порядка ограниченного подмножества вершин, наблюдаемых в процессе мониторинга поведения.

Для оценки адекватности модели в качестве нулевой гипотезы выбрано утверждение о том, что поведение пользователя является нормальным. Для этого загружены лог-записи десяти пользователей ERP-системы 1С:Розница. Посторонние записи получены при выполнении пользователями инструкции-сценария, представляющей случай поведенческой аномалии.

Для наблюдения за пользователями и получения лог-записей (табл. 1) использовано расширение для веб-браузера Google Chrome. В качестве идентификаторов элементов интерфейса пользователя использованы значения атрибута xpath. Малоиспользуемые элементы интерфейса (например, отдельные строки таблиц) объединены в один узел на основе иерархии xpath. В роли характеристик текстовых данных использованы именованные сущности, префиксы телефонных номеров и подмножества почтовых индексов. Для выявления именованных сущностей использована библиотека pymorphy3.

Таблица 1

Характеристики набора данных

Пример последовательности действий	Тип поведения
Количество лог-записей	302,3 шт./ч
Количество элементов интерфейса	18
Количество вершин графа де Брюина второго порядка	324
Количество элементов ограниченного множества вершин графа де Брюина второго порядка	49
Размер временного окна	24
Количество признаков текстовых данных	40 (32 – именованные сущности, 8 – географические коды)

При выполнении экспериментального анализа проверены два сценария утечек данных:

1. Единичные случаи (табл. 2). Пользователь в течение рабочего дня просматривает несколько интересующих записей. Содержание записей соответствует служебным обязанностям сотрудника, но записи просмотрены с целью хищения данных. Последовательность работы с инструментами КИС при обычном исполнении бизнес-процессов нарушена. Например, пользователь совершает переходы: главная страница, новое обращение, список клиентов, список клиентов, анкета клиента, анкета клиента. При этом отсутствуют узлы графа «оформление обращения», «обращение создано».

2. Выгрузка записей несвойственной тематики. Пользователь в течение рабочего дня находит интересующие записи, открывая справочник с анкетами клиентов. Содержание записей не соответствует обычной работе пользователя. Например, пользователь открывает анкеты клиентов с несвойственным географическим кодом.

Таблица 2

Примеры обычной и аномальной последовательности действий для сценария № 1

Пример последовательности действий	Тип поведения
(0) Главная страница (1) Заказы (2) Новый (3) Контрагенты (4) Выбор (5) Продукция ….. (7) Продукция (8) Оформить (9) Главная страница	Обычное
(0) Главная страница (1) Заказы (2) Новый (3) Контрагенты ..... (8) Контрагенты (9) Главная страница	Аномальное (выгрузка данных контрагентов)

Для имитации поведенческих аномалий в отдельные временные окна добавлены лог-записи, полученные при наблюдении за пользователями при выполнении сценариев № 1 и № 2.

Разреженность многомерных данных, представленных коллекцией векторов временных окон, накладывает ограничения на выбор алгоритма для поиска аномалий, включая алгоритмы на основе сравнения гистограмм [14]. В данной рассматриваемой модели отдельные лог-записи или векторы временных окон могут иметь менее 50% ненулевых атрибутов. Для поиска аномальных значений использована нейронная сеть архитектуры «автокодировщик» [17] с тремя скрытыми слоями и функцией активации сигмоида.

Произведена проверка идентификации поведенческих аномалий с помощью автокодировщика с применением данных графа первого и второго порядков для сценария № 1 (табл. 3) и сценария № 2 (табл. 4). При нарушении исполнения бизнес-процессов возникало превышение ошибки репликации (рис. 2). Использование графа де Брюина второго порядка позволяет повысить точность определения аномалий.

Таблица 3

Результаты эксперимента по поиску посторонних элементов сценария № 1

Алгоритм	Площадь под ROC-кривой
Автокодировщик вершин графа второго порядка	0,909
Автокодировщик на основе текстовых данных и вершин графа второго порядка	0,813
Автокодировщик на основе текстовых данных и вершин графа первого порядка	0,631

Таблица 4

Результаты эксперимента по поиску посторонних элементов сценария №2

Алгоритм	Площадь под ROC-кривой
Автокодировщик на основе текстовых данных	0,874
Автокодировщик на основе текстовых данных и вершин графа второго порядка	0,821

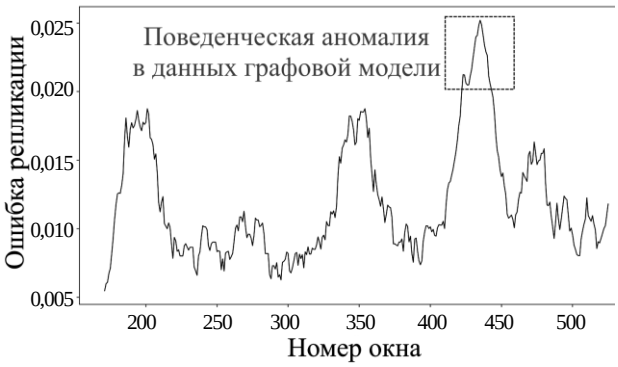

Рис. 2. Поведенческая аномалия для сценария № 1

Поиск аномалий в отдельном подмножестве признаков документной модели (текстовых данных) более предпочтителен (см. табл. 4).

Заключение

В результате выполненной работы создана модель поведения пользователя корпоративной информационной системы, которая позволяет находить поведенческие аномалии, характеризующиеся изменениями в последовательностях действий пользователя или использованием текстовых данных нетипичной тематики. Использование поведенческого графа второго порядка позволяет повысить точность классификации при использовании автокодировщика для поиска посторонних значений. С использованием модели обнаруженная аномалия может быть детализирована в виде подграфа поведения с указанием последовательностей смены состояний клиента КИС и связанными с изменением состояний признаков текстовых данных.

Работа выполнена в рамках проекта «Гранты ИБ МТУСИ» 2022.

Литература

1. Shabtai A. A survey of data leak-age detection and prevention Solutions. / A. Shabtai, Y. Elovici, L. Rokach. – Berlin, Germany: Springer, 2012. – 92 p.

2. Исакова Т. На работу, как на фишинг / Т. Исакова, Н. Королев // Газета «Коммерсантъ». – 08.08.2022. – Ст. 18.

3. «Ростелеком» оштрафовали на 60 000 рублей за утечку пользовательских данных // Хакер [Электронный ресурс]. – Режим доступа: URL: https://xakep.ru/2023/04/19/rostelekom-penalty/ (дата обращения: 19.04.2023).

4. Курашева А. Количество утечек данных в крупных компаниях выросло в 1,5 раза // Ведомости. – 2023. – 12.05. – Ст. 19.

5. Принцип работы Solar Dozor // Ростелеком [Электронный ресурс]. – Режим доступа: URL: https://rt-solar.ru/products/solar_dozor/architecture/ (дата обращения: 24.09.2023).

6. Евсиков В. Пробей меня полностью! Кто, как и за сколько пробивает персональные данные в России // Хакер. – 2020. – № 10 [Электронный ресурс]. – Режим доступа: https://xakep.ru/2020/10/09/personal-data/ (дата обращения: 05.09.2023).

7. Denning D.E. An Intrusion Detection Model // IEEE transactions on software engineering. – 1987. – Vol. SE-13. – P. 222–232.

8. Hu J. Anomalous User Activity Detection in Enterprise Multi-Source Logs / J. Hu, T. Baoming, D. Lin // International Conference on Data Mining Workshops. – New Orleans, USA: IEEE, 2017. – P. 797–803.

9. Mohotti W. Efficient outlier detection in text corpus using rare frequency and ranking / W. Mohotti, R. Nayak // ACM Transactions on Knowledge Discovery from Data. – 2020. – Vol. 14, No. 6. – P. 1–30.

10. Haixuan G. LogBERT: Log Anomaly Detection via BERT / G. Haixuan, Y. Shuhan, W. Xintao// International Joint Conference on Neural Networks (IJCNN). – Shenzhen, China: IEEE, 2021. – P. 1–8.

11. Weibin M. et al. LogAnomaly: Unsupervised Detection of Sequential and Quantitative Anomalies in Unstructured Logs // Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. – Macao, China: IJCAI, 2019. – P. 4739–4745.

12. A Graph Embedding Approach to User Behavior Anomaly Detection / A. Modell, J. Larson, M. Turcotte, A. Bertiger // GTA 2.0: The 5th IEEE Big Data Workshop on Graph

Techniques for Adversarial Activity Analytics. – New Jersey, USA: IEEE, 2021. – P. 2650–2655.

13. Modelling User Behavior Dynamics with Embeddings / L. Han, A. Checco, D. Difallah, G. Demartini, S. Sadiq // CIKM '20: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. – NY, USA: ACM, 2020. – P. 445–454.

14. Boniol P. Series2Graph: graph-based subsequence anomaly detection for time series / P. Boniol, T. Palpanas // Proceedings of the VLDB Endowment. – Tokyo: VLDB Endowment, 2020. – P. 1821–1834.

15. Bruijn de N.G. A combinatorial problem // Proceedings of the Section of Sciences of the Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam. – 1941. – Vol. 49, No. 7. – P. 758–764.

16. Pevny T. Loda: Lightweight on-line detector of anomalies // Machine Learning. – 2016. – Vol. 102. – P. 275–304.

17. Autoencoder-based outlier detection for sparse, high dimensional data / W. Chen, H. Li, H. Li, A. Arshad // Proceedings of 2020 IEEE International Conference on Big Data (Big Data). – Atlanta, USA: IEEE, 2020. – P. 2735–2742.

Банокин Павел Иванович
Преп. каф. комплексной информационной безопасности электронно-вычислительных систем (КИБЭВС)
Томского государственного ун-та
систем управления и радиоэлектроники (ТУСУР)
Ленина пр-т, 40, г. Томск, Россия, 634050
Тел.: + 7 (382-2) 70-15-29
Эл. почта: pavel805@gmail.com

Banokin P.I.
Model of corporate information system user behavior

A behavior model of corporate user is presented. The model is based on a De Bruijn graph, where the vertices store the features of text data. The values of textual features are extracted with regard to the data displayed in the client application information system. The model could be used to identify the behavioral anomalies, that could be linked to internal data leaks. The scope of detected behavioral anomalies includes scenarios of the corrupted business process flow and the data leakage of an unusual topic. The model is evaluated by criteria of efficiency, adequacy and completeness.

Keywords: internal data leaks, corporate information systems, behavior model.
DOI: 10.21293/1818-0442-2023-26-4-78-83

References

1. Shabtai A., Elovici Y., Rokach L. *A survey of data leakage detection and prevention Solutions*. Berlin, Germany, Springer, 2012, 92 p., pp. 39-46,

2. Isakova T., Korolev N. *Na rabotu kak na fishing*. [To go to work as to go fishing]. Newspaper «Kommersant», 08.08.2022, art. 18 (in Russ.).

3. *«Rostelekom» oshtrafovali na 60 000 rublei za utechku polzovatelskih dannyh* [Rostelecom is fined 60 000 rubles for personal data leakage]. *Haker* [Hacker]. Available at: https://xakep.ru/2023/04/19/rostelekom-penalty/, free (Accessed: April 19, 2023). (in Russ.).

4. Kurasheva A. *Kolichestvo utechek dannyh v krupnih kompaniyah uvelichlos v 1.5 raza* [The number of data leaks increased 1.5 times in large corporations]. *Vedomosti*, 2023, 12.05, article 19 (in Russ.).

*5. Printsip raboty Solar Dozor [The princi*ple of Solar Dozor]. Rostelecom. URL: https://rt-solar.ru/products/solar_dozor/architecture/ (Accessed: September 24, 2023) (in Russ.).

6. Evsikov V. *Probej menya polnost'yu. Kto, kak I za skolko probivaet personalnie dannie v Rossii* [Get my full personal data details! Who, how and for how much does it cost to access personal data in Russia?]. *Haker* [Hacker], 2020, no. 10. Available at: https://xakep.ru/2020/10/09/personal-data/ (Accessed: September 5, 2023) (in Russ.).

7. Denning D.E. An Intrusion Detection Model. *IEEE transactions on software engineering*, 1987, vol. SE-13, pp. 222–232.

8. Hu J., Baoming T., Lin D. Anomalous User Activity Detection in Enterprise Multi-Source Logs. *International Conference on Data Mining Workshops*, New Orleans, USA, IEEE, 2017, pp. 797–803.

9. Mohotti W., Nayak R. Efficient outlier detection in text corpus using rare frequency and ranking. *ACM Transactions on Knowledge Discovery from Data*, 2020, vol. 14, no. 6, pp. 1–30.

10. Haixuan G., Shuhan Y., Xintao W. LogBERT: Log Anomaly Detection via BERT. *International Joint Conference on Neural Networks* (*IJCNN*), Shenzhen, China, IEEE, 2021, pp. 1–8.

11. Weibin M. [et al.] LogAnomaly: Unsupervised Detection of Sequential and Quantitative Anomalies in Unstructured Logs. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, Macao, China, IJCAI, 2019, pp. 4739–4745.

12. Modell A., Larson J., Turcotte M., Bertiger A. A Graph Embedding Approach to User Behavior Anomaly Detection. GTA 2.0: *The 5th IEEE Big Data Workshop on Graph Techniques for Adversarial Activity Analytics*, New Jersey, USA, IEEE, 2021, pp. 2650–2655.

13. Boniol P., Palpanas T. Series2Graph: graph-based subsequence anomaly detection for time series. *Proceedings of the VLDB Endowment*, Tokyo, Japan, VLDB Endowment, 2020, pp. 1821–1834.

14. Han L., Checco A., Difallah D., Demartini G., Sadiq S. Modelling User Behavior Dynamics with Embeddings. CIKM '20: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, NY, USA, ACM, 2020, pp. 445–454.

15. Bruijn de N.G. A combinatorial problem. *Proceedings of the Section of Sciences of the Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam*, 1941, vol. 49, no. 7, pp. 758–764.

16. Pevny T. Loda: Lightweight on-line detector of anomalies. *Machine Learning*, 2016, vol. 102, pp. 275–304.

17. Chen W., Li H., Li H., Arshad A. Autoencoder-based outlier detection for sparse, high dimensional data. *Proceedings of 2020 IEEE International Conference on Big Data* (*Big Data*), Atlanta, USA: IEEE, 2020, pp. 2735–2742.

Pavel I. Banokin
Lecturer, Department of Complex Information
Security of Electronic Computer Systems (KIBEVS),
Tomsk State University of Control Systems
and Radioelectronics (TUSUR)
40, Lenin pr., Tomsk, Russia, 634050
Phone: + 7 (382-2) 70-15-29
Email: pavel805@gmail.com