Московский государственный университет
имени М.В.Ломоносова

На правах рукописи

Царёв Дмитрий Владимирович

МЕТОДЫ И ПРОГРАММНЫЕ СРЕДСТВА АНАЛИЗА ПОВЕДЕНИЯ ПОЛЬЗОВАТЕЛЕЙ ПРИ РАБОТЕ С ТЕКСТОВЫМИ ДАННЫМИ ДЛЯ РЕШЕНИЯ ЗАДАЧ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

Специальность 05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Автореферат диссертации на соискание ученой степени
кандидата физико-математических наук

Москва – 2017

Работа выполнена на кафедре автоматизации систем вычислительных комплексов факультета вычислительной математики и кибернетики Московского государственного университета имени М.В. Ломоносова.

Научные руководители: доктор физико-математических наук,
профессор Машечкин Игорь Валерьевич

кандидат физико-математических наук,
доцент Петровский Михаил Игоревич

Официальные оппоненты: Безродный Борис Федорович,
доктор технических наук, профессор,
ОАО «НИИАС»,
заместитель руководителя Центра кибербезопасности

Крейнес Михаил Григорьевич,
кандидат физико-математических наук,
ООО «БАЗИСНЫЕ ТЕХНОЛОГИИ»,
генеральный директор

Ведущая организация: Межведомственный суперкомпьютерный центр Российской академии наук – филиал Федерального государственного учреждения «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук» (МСЦ РАН – филиал ФГУ ФНЦ НИИСИ РАН)

Защита диссертации состоится 20 июня 2017 г. в 11:00 на заседании диссертационного совета Д 501.001.44 при Московском государственном университете имени М.В. Ломоносова по адресу: 119991, ГСП-1, Москва, Ленинские горы, МГУ, 2-й учебный корпус, факультет вычислительной математики и кибернетики, аудитория 685.

С диссертацией можно ознакомиться в Научной библиотеке Московского государственного университета имени М.В. Ломоносова по адресу: 119192, г. Москва, Ломоносовский проспект, д. 27, а также на официальном сайте факультета ВМК МГУ http://www.cmc.msu.ru в разделе «Диссертации».

Автореферат разослан «___» апреля 2017 г.

Ученый секретарь диссертационного совета
доктор физико-математических наук, доцент Шестаков О.В.

Общая характеристика работы

Актуальность темы исследования

Утечка данных является одной из самых опасных угроз для современных компаний. Для предотвращения утечек традиционно применяются системы класса DLP (англ. *Data Loss Prevention*), позволяющие обнаруживать конфиденциальную информацию в потоках данных, покидающих информационный периметр организации. Однако всё больше экспертов сходится во мнении, что использование DLP-систем недостаточно эффективно, и поэтому утечки необходимо определять ещё до стадии пересылки данных за информационный периметр. Данное утверждение основывается на исследованиях[1], показывающих, что от момента, когда пользователь решает украсть данные до непосредственно пересылки данных, проходит от нескольких недель до нескольких месяцев, которые уходят на стадию подготовки утечки. В данной стадии поведение пользователя отличается от его обычной легитимной активности как по набору выполняемых действий, так и по содержанию обрабатываемой информации. Поэтому за последние несколько лет активное развитие получило направление анализа поведения пользователей с целью обнаружения аномалий. Аномальное поведение может свидетельствовать о том, что пользователь не является тем, от имени кого он авторизовался (задача идентификации пользователей), или пользователь интересуется корпоративными документами, которые не относятся к его текущей рабочей деятельности, что является признаком потенциальной утечки информации (задача раннего обнаружения попыток хищения информации).

В настоящий момент сформировался самостоятельный класс систем информационной безопасности, в основе которых лежат методы машинного обучения для выявления признаков несвойственного (аномального) поведения пользователей. Компания Gartner[2] данный класс систем обозначает как UEBA (англ. *User and Entity Behavior Analytics* — анализ поведения пользователей и систем). UEBA-системы, в отличие от DLP, осуществляют мониторинг широкого спектра действий пользователя и принимают решения не на основе экспертно сформированных политик безопасности, а на основе исторических данных о легитимной работе пользователя. Данные системы обнаруживают ранние признаки утечки, поэтому их основная цель состоит не в блокировке действий пользователей, а в предоставлении аналитических данных службе ИБ с описанием того, почему выявленные действия являются

[1] ObserveIT Data Loss Prevention Capabilities [Электронный ресурс]. — Электрон. дан. — [Б. м.] : ObserveIT, 2015. — Режим доступа: http://www.observeit.com/blog/observeit-data-loss-prevention-capabilities-1.

[2] Gartner, Inc.— компания, специализирующаяся на рынках информационных технологий, является мировым лидером в области исследований и консалтинговых услуг (http://www.gartner.com/technology/about.jsp).

аномальными для конкретного пользователя. UEBA-системы на основе методов машинного обучения выполняют построение и применение моделей поведения (профилей) пользователей с целью выявления признаков аномального поведения.

Степень разработанности темы

Обычно целью внутренних вторжений является получение доступа к текстовой информации (отчёты, договоры, техническая документация, электронная почта и т.п.), поэтому ключевым является выявление аномального поведения пользователей при работе с текстовыми данными. Существующие UEBA-системы с помощью методов машинного обучения анализируют данные об операциях пользователя (контекстную информацию), которые являются хорошо структурированными, например, данные системных журналов ОС, журналов SIEM, IDS/IPS, DLP-систем; данные об операциях с файлами, электронной почтой. Анализ содержимого обрабатываемых пользователем текстовых данных представляет более сложную задачу и не рассматривается в существующих решениях UEBA-систем. Во-первых, текст является неструктурированной информацией, а во-вторых, представляет данные гораздо большего объёма, зачастую содержащие информационный шум. Поэтому на сегодняшний день существующие подходы не способны выявить случаи нелегитимной активности пользователя при характерных для него действиях, но с нелегальным содержимым (контентом). Кроме того, только лишь анализ структурированной информации об операциях пользователя не даёт стопроцентную точность обнаружения утечки.

Новизну и актуальность выбранной темы диссертации подтверждает отчёт Gartner[3], в котором также подчёркивается, что анализ текстовых данных является гораздо более сложной задачей, чем анализ структурированных данных об операциях. Поэтому Gartner ожидает появление данного функционала в UEBA-системах в течение следующих нескольких лет, отмечая при этом важность анализа пользовательской текстовой информации для понимания и оценки действий пользователя.

Цели и задачи

Целью диссертационной работы является исследование и разработка математического и программного обеспечения обнаружения аномального поведения пользователей на основе анализа содержимого потока обрабатываемых текстовых данных с использованием методов машинного обучения для задач информационной безопасности.

Объектом исследования диссертационной работы является поведенческая информация пользователей при работе с электронными текстовыми документами. Под поведенческой

[3] Gartner. Market Guide for User and Entity Behavior Analytics [Электронный ресурс]. — Электрон. дан. — [Б. м.] : Gartner, 2015. — Режим доступа: https://www.gartner.com/doc/reprints?id=1-2NK6M1R&ct=150922&st=sb.

информацией пользователя будем понимать данные об операциях, выполняемых пользователем с электронными документами, и данные о содержимом этих документов.

Для достижения поставленной цели необходимо решение следующих задач:

1. Разработать модель представления поведенческой информации пользователя о его работе с текстовыми данными и исследовать возможность применения методов удаления информационного шума.
2. Разработать методы обнаружения аномального поведения пользователя при работе с текстовыми данными, используя выбранную модель представления поведенческой информации. Разработанные методы должны быть основаны на машинном обучении и служить для построения и применения индивидуальных моделей поведения пользователей.
3. Разработать архитектуру и реализовать экспериментальный образец программного комплекса, выполняющего сбор поведенческой информации, построение и применение индивидуальных моделей поведения пользователей на основе разработанного комплекса алгоритмов для обнаружения аномального поведения.

Научная новизна заключается в предложенном новом подходе к анализу и моделированию поведения пользователя, основанном на отображении содержимого потока электронных документов в тематическое пространство, формируемое с использованием неотрицательной матричной факторизации. Изменение значений весов тематик во времени представляет многомерный временной ряд, описывающий историю поведения пользователя при работе с текстовыми данными. Анализ такого временного ряда позволяет определять факты аномального поведения пользователя. Разработаны новые методы, основанные на расчёте оценки принадлежности документов пользователя к характерным для него тематикам, и методы оценки отклонения тематической направленности пользователя от спрогнозированных значений.

Практическая значимость работы состоит в разработке и реализации экспериментального образца программного комплекса обнаружения аномального поведения пользователей по особенностям работы с текстовой информацией, предназначенного для решения задач информационной безопасности. Полученные результаты диссертационной работы могут послужить основой для построения перспективных современных систем информационной безопасности класса UEBA, которые будут включать средства анализа содержимого обрабатываемых пользователями текстовых данных. Причём могут использоваться как все разработанные программные модули для осуществления сбора поведенческой информации, построения и применения индивидуальных моделей поведения

пользователей, так и только модули, служащие для сбора и представления в структурированном виде содержимого обрабатываемых пользователями текстовых данных.

Методология и методы исследования

При получении основных результатов диссертации использовались методы теории машинного обучения и анализа текстов на естественном языке, а также проведённые экспериментальные исследования на примере набора реальной корпоративной электронной почты. При разработке программных модулей экспериментальной системы обнаружения аномального поведения пользователей по особенностям работы с текстовой информацией использовались методы объектно-ориентированного анализа и проектирования.

Личный вклад автора заключается в выполнении основного объема теоретических и экспериментальных исследований, а также в разработке архитектуры и реализации экспериментального образца мультиагентного программного комплекса обнаружения аномального поведения пользователей по особенностям работы с текстовой информацией. Автор выполнил анализ и оформление полученных результатов диссертационной работы в виде публикаций, научных докладов, патента на полезную модель и двух свидетельств о государственной регистрации программ для ЭВМ.

Степень достоверности и апробация результатов

Результаты, представленные в работе, докладывались: весной 2016 года на научном семинаре Института системного программирования РАН «Управление данными и информационные системы» под руководством академика РАН В.П. Иванникова; осенью 2016 года на научном семинаре кафедры автоматизации систем вычислительных комплексов факультета вычислительной математики и кибернетики Московского государственного университета имени М.В. Ломоносова.

Также результаты диссертационной работы докладывались автором на следующих конференциях:

- Научная конференция «Ломоносовские чтения» (Россия, Москва, 2011).
- 11th International Conference on Hybrid Intelligent Systems (Малайзия, Малакка, 2011).
- Научная конференция «Тихоновские чтения» (Россия, Москва, 2012).
- Научная конференция «Ломоносовские чтения» (Россия, Москва, 2012).
- 16-я Всероссийская конференция «Математические методы распознавания образов» (Россия, Казань, 2013).
- 14th International Conference on Hybrid Intelligent Systems (Кувейт, 2014).
- Научная конференция «Тихоновские чтения» (Россия, Москва, 2015).

Основные результаты по теме диссертации изложены в 21 публикации, 12 из которых изданы в журналах, рекомендованных ВАК: перечень ВАК — 6; система цитирования

Scopus — 6 (из них 2 входят в систему цитирования Web of Science). Статьи [3, 13, 16] являются переводом на английский язык статей [2, 12, 15] соответственно.

Результаты данной работы использовались в следующих НИР:

- «Разработка программного комплекса мониторинга и анализа работы пользователей с документами в корпоративных сетях» (проект РФФИ № 12-07-00585), 2012-2014 гг.
- «Разработка вычислительных методов объективной оценки качества научно-технических документов на естественных языках» (Государственный контракт № 14.514.11.4016), 2012-2013 гг.
- «Исследование и разработка инновационной технологии построения программных средств обеспечения компьютерной безопасности, основанных на использовании методов машинного обучения и математической статистики для анализа данных поведенческой биометрии пользователей при работе в рамках стандартного человеко-машинного интерфейса, для решения задач активной аутентификации и идентификации пользователей, обнаружения внутренних вторжений и предотвращения попыток хищения конфиденциальной информации» (Работы выполнены при финансовой поддержке Минобрнауки России. Соглашение № 14.604.21.0056 о предоставлении субсидии. Уникальный идентификатор прикладных научных исследований RFMEFI60414X0056), 2014-2016 гг.
- Грант РФФИ № 16-29-09555\16 по направлению «Безопасность и противодействие терроризму», 2016-2018 гг.

Структура и объём диссертации

Диссертация состоит из введения, четырёх глав, заключения и списка литературы. Полный объём диссертации составляет 143 страницы. Список литературы содержит 137 наименований.

Краткое содержание диссертации

Во **введении** показывается актуальность темы исследования и степень её разработанности на примере сравнения функционала традиционных DLP-систем, применяемых для предотвращения утечек данных, и активно развивающихся в последние несколько лет UEBA-систем, в основе которых лежат методы машинного обучения для выявления признаков аномального поведения пользователей. Далее ставится цель и формулируются задачи работы, обосновываются научная новизна, практическая и теоретическая значимость их решения. В заключение приводится краткий обзор содержания диссертации.

Первая глава посвящена исследованию существующих подходов к анализу текстовой информации, применяемых в современных программных системах, функционал которых направлен на управление контентной информацией организации. К данным системам были отнесены системы следующих классов: системы управления корпоративным контентом (англ. *Enterprise Content Management*, *ECM*), которые также включают средства электронного раскрытия информации (англ. *eDiscovery*); DLP-системы предотвращения утечек данных. Обзор систем данных классов приводится соответственно в **подразделе 1.1** и **подразделе 1.2**.

На основе проведённого аналитического обзора формулируются направления дальнейших исследований:

- *Выбор модели представления поведенческой информации.* Исследовать возможность описания текстового контента пользователя с помощью характерных для него последовательностей семантически связанных слов — тематик.
- *Выбор методов обнаружения аномалий.* Исследовать возможность применения методов прогнозирования временных рядов для определения интервалов времени аномальной работы пользователя с текстовыми документами. Исследовать возможность применения методов классификации на основе машинного обучения для определения фактов работы пользователя с несвойственными ему текстовыми документами.

Во **второй главе** проводится исследование и разработка модели представления поведения пользователя на основе анализа содержимого потока обрабатываемых текстовых данных. Формально поток документов пользователя представляет множество $x = \{(d_1, t_1), \ldots, (d_n, t_n)\} \subset \mathrm{X} = \mathrm{D} \times \mathrm{T}$, каждый элемент которого представляет пару (d_j, t_j) — анализируемый объект $(1 \le j \le n)$, где d_j — документ, содержащий текстовые данные пользователя, t_j — временная метка, соответствующая обращению пользователя к d_j. Для формирования модели поведения пользователя по потоку необходимо описывать его текстовые данные набором признаков, изменения значений которых будут определять поведение пользователя с течением времени. Коллекцию документов $C = (d_1, \ldots, d_n)$ требуется представить в виде числовой матрицы $A \in \mathbb{R}^{m \times n}$, строки которой соответствуют признакам, а столбцы — документам. Каждый документ d_j $(1 \le j \le n)$ представляется в виде числового вектора $A_j = [a_{1,j}, a_{2,j}, \ldots, a_{m,j}]^T$, фиксированной размерности m, где m — число признаков коллекции документов, а i-ая $(1 \le i \le m)$ компонента вектора A_j определяет вес i-го признака в документе d_j. Тогда матрица A задаёт модель поведения пользователя в виде m-мерного временного ряда, показывающего изменение весов соответствующих признаков в пользовательском потоке.

В **подразделе 2.1** рассмотрена классическая модель представления текста в векторном виде «мешок слов» (англ. «*bag-of-words*»). В качестве признаков в данной модели

используются термы — лексемы, входящие в текст. Основными недостатками данной модели представления, применённой к коллекции документов, являются высокая размерность пространства признаков (термов), и игнорирование семантических взаимосвязей между словами в получаемых признаках. Как следствие, формируемая матрица A содержит большое число строк и является сильно разреженной. Что, в свою очередь, приводит к низкой скорости работы алгоритмов обнаружения аномалий и к невозможности применения методов прогнозирования временных рядов для строк матрицы A.

Подраздел 2.2 посвящён исследованию тематических моделей для представления пользовательских текстовых данных с помощью характерных для него тематик. Использование таких моделей представления документов приводит к уменьшению пространства признаков за счёт объединения разных, но семантически связанных, термов в один признак — тематику. В настоящей работе для тематического моделирования используется один из наиболее перспективных[4] на сегодняшний день подходов, основанный на применении неотрицательной матричной факторизации в латентно-семантическом анализе.

В общем случае латентно-семантический анализ работает с матричным представлением коллекции документов $C = (d_1, \ldots, d_n)$, получаемым с помощью модели «мешок слов»: $A \in \mathbb{R}^{m \times n}$, где m — число различных термов коллекции, а n — число документов. Определение основных тематик коллекции и представление текстов в пространстве тематик осуществляется применением к матрице A одного из матричных разложений, например: сингулярного разложения, неотрицательной матричной факторизации.

Цель неотрицательной матричной факторизации, применённой к матрице $A \in \mathbb{R}^{m \times n}$, состоит в нахождении матриц $W_k \in \mathbb{R}^{m \times k}$ и $H_k \in \mathbb{R}^{k \times n}$ с неотрицательными элементами, которые минимизируют целевую функцию: $f(W_k, H_k) = \frac{1}{2}\|A - W_k H_k\|_F^2 + \frac{\alpha}{2}\|W_k^T W_k - I\|_F^2$, где $k << min(m, n)$, $\alpha \geq 0$ — параметр ортонормированности W_k, если задать $\alpha > 0$, то накладывается дополнительное условие: $W_k^T \cdot W_k = I$.

Формально выбранная тематическая модель коллекции текстовых документов $C = (d_1, \ldots, d_n)$ представляет совокупность (L_m, W_k, H_k), где: L_m — словарь, содержащий m термов коллекции C; $W_k \in \mathbb{R}^{m \times k}$ — матрица отображения между пространством k тематик и пространством m термов; $H_k = [H^1, \ldots, H^n] \in \mathbb{R}^{k \times n}$ — матрица представления документов в пространстве тематик. Под тематическим пространством далее будем понимать совокупность (L_m, W_k).

[4] Kuang D., Choo J., Park H. Nonnegative Matrix Factorization for Interactive Topic Modeling and Document Clustering // Partitional Clustering Algorithms. — Springer International Publishing, 2015. — С. 215-243.

В **подразделе 2.3** приводится описание предложенной тематической модели поведения пользователя. Формально предлагаемая модель поведения, сформированная по потоку пользовательских текстовых документов $x = \{(d_1, t_1), \ldots, (d_n, t_n)\}$, представляет совокупность (L_m, W_k, H_k, T_n), где: (L_m, W_k, H_k) — тематическая модель коллекции документов $C = (d_1, \ldots, d_n)$, основанная на ортонормированной неотрицательной матричной факторизации; $T_n = (t_1, \ldots, t_n)$ — временные метки каждого документа (см. Рисунок 1). Столбец H^j $(1 \le j \le n)$ матрицы H_k соответствует тематической направленности пользователя во время t_j.

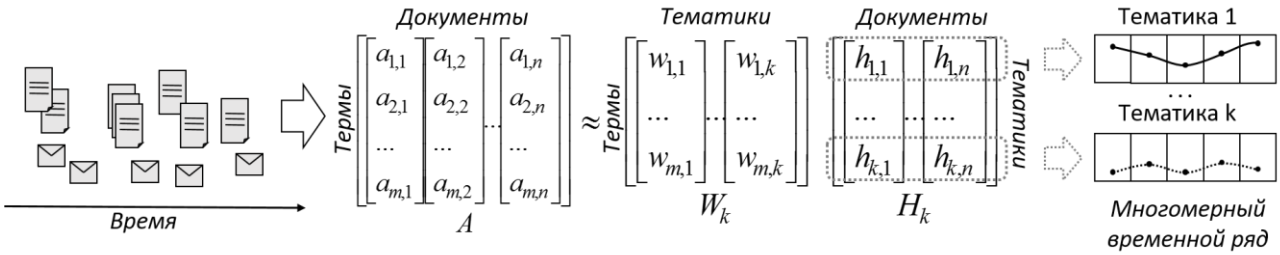

Рисунок 1 — Формирование тематической модели поведения пользователя.

Наличие свойства ортонормированности W_k является необходимым, т.к. поведенческая модель может применяться для анализа дальнейшей активности пользователя $y = \{(d_{new}, t_{new})\}$, где $d_{new} \notin C$, $t_{new} > t_n$ Поэтому необходимо представлять новые документы $d_{new} \notin C$ в уже сформированном тематическом пространстве (L_m, W_k): $W_k^T \cdot A_{new} = H_{k\_new}$, где A_{new} — вектор представления d_{new} в модели «мешок слов» со словарём термов L_m.

Подраздел 2.4 посвящён задаче удаления информационного шума из документов, которая заключается в оценке значимости (релевантности) отдельных фрагментов текста и последующего удаления из результирующего документа наименее значимых фрагментов, при этом оставляемые фрагменты должны описывать все главные темы исходного текста. Таким образом, удаление информационного шума из документов может привести к уменьшению объёма обрабатываемых данных и к повышению точности работы алгоритмов обнаружения аномалий.

Рассматриваемая задача тесно связана с задачей автоматического аннотирования, для решения которой также необходимо сформировать аннотацию, состоящую из наиболее значимых фрагментов текста. На сегодняшний день наиболее популярные методы автоматического аннотирования, которые вычисляют релевантность фрагментов текста, основаны на тематическом моделировании текстов с использованием латентно-семантического анализа, применённого к коллекции отдельных фрагментов (например, предложений) анализируемого документа.

Формально требуется построить функцию ранжирования R: $\mathbb{R}^{m \times n} \to \mathbb{R}^n$, которая по заданному матричному представлению фрагментов текста $A \in \mathbb{R}^{m \times n}$ вычисляет степень

релевантности каждого фрагмента. Релевантность фрагмента представляет агрегированную оценку соответствия данного фрагмента ключевым тематикам текста и прямо зависит от весов тематик во фрагменте. Чем больше веса тематик во фрагменте, тем больше его релевантность.

В рамках данного подраздела представлен новый метод вычисления релевантности фрагментов текста, основанный на оценке весов тематик в нормализированном пространстве тематик, получаемом с помощью неотрицательной матричной факторизации.

Неотрицательная матричная факторизация даёт не единственное решение, поэтому первый этап предложенного метода вычисления релевантности фрагментов текста состоит в нормировке пространства k тематик: $A_k = W_k \cdot H_k = Wn_k \cdot Hn_k$, где $Wn_k = W_k \cdot diag(\|w^1\|^{-1}, \ldots, \|w^k\|^{-1})$, $Hn_k = diag(\|w^1\|, \ldots, \|w^k\|) \cdot H_k$, , $\|w^l\| = \sqrt{\sum_{p=1}^{m} w_{pl}^2}$, $1 \le l \le k$.

Второй этап заключается в оценке глобальных весов полученных тематик во всём документе. Столбцы матрицы $Hn_k = [hn_{ij}]$ соответствуют n фрагментам документа в нормированном пространстве k тематик. Каждая из k строк Hn_k соответствует вектору, показывающему насколько сильно представлена соответствующая тематика в каждом из n фрагментов. Тем самым, чем больше длина вектор-строки матрицы Hn_k, тем соответствующая тематика «больше» представлена во всем документе. Исходя из этого, глобальный вес тематики l оценивается как длина l-ой вектор-строки матрицы Hn_k: $\|hn_l\| = \sqrt{\sum_{q=1}^{n} hn_{lq}^2} = \|w^l\| \cdot \sqrt{\sum_{q=1}^{n} h_{lq}^2} = \|w^l\| \cdot \|h_l\|$, $1 \le l \le k$. Тогда релевантность j-ого фрагмента вычисляется как норма вектора, являющегося результатом поэлементного умножение вектора глобальных весов тематик и вектора весов тематик в рассматриваемом фрагменте:

$$R_j = \sum_{i=1}^{k} \left((\|w^i\| \cdot \|h_i\|) \cdot \left(\|w^i\| \cdot h_{ij} \right) \right) = \sum_{i=1}^{k} \left(\|w^i\|^2 \cdot \|h_i\| \cdot h_{ij} \right).$$

Для оценки качества методов вычисления релевантности фрагментов текста была рассмотрена задача автоматического аннотирования, которая заключалась в составлении аннотаций из наиболее релевантных предложений документа. Эксперименты, проведённые на эталонном наборе данных DUC 2002[5] с использованием стандартных метрик оценки качества аннотаций ROUGE, показали превосходство предложенного автором метода.

Третья глава посвящена исследованию и разработке методов машинного обучения для обнаружения аномального поведения пользователя при работе с текстовыми данными. Более формально задачу обнаружения аномального поведения пользователя можно сформулировать следующим образом: по заданному потоку текстовых документов $x = \{(d, t)\} \subset X$ требуется

[5] DUC 2002 Guidelines [Электронный ресурс]. — Электрон. дан. — [Б. м.] : NIST, 2014. — Режим доступа: http://www-nlpir.nist.gov/projects/duc/guidelines/2002.html.

построить функцию f: $(d, t) \rightarrow \mathbb{R}$, называемую решающей функцией, такую, что для анализируемого объекта $(d, t) \in X$ ставится в соответствие значение аномальности $a \in \mathbb{R}$, которое зависит от того, насколько объект (d, t) «похож» на элементы множества x.

Корпоративный пользователь за относительно длительные промежутки времени (например, 12, 24 часа, рабочее/нерабочее время) обычно успевает интересоваться всем характерным для себя текстовым контентом — рабочие документы, новости и т.п. Однако, если рассмотреть короткие интервалы времени (например, 15, 30 минут), то очерёдность обращения пользователя к документам определённых тематик трудно предугадать. Исходя из указанной специфики было предложено два подхода к обнаружению аномального поведения пользователя:

1. Прогнозирование тематической направленности пользователя по «длительным» интервалам времени.
2. Оценка принадлежности документа, с которым работает пользователь, к характерным тематикам анализируемого пользователя.

Первый подход позволит оценивать общую аномальность поведения пользователя за «длительное» время, а второй необходим для оценки аномальности каждого обращения пользователя к документам.

В **подразделе 3.1** описывается сложившаяся на сегодняшний день в научных статьях практика проведения экспериментальных исследований в области обнаружения внутренних угроз. Формулируются требования к точности разрабатываемых методов обнаружения аномального поведения пользователей. После чего приводится описание базового сценария экспериментальных исследований для верификации разрабатываемых методов.

Для проведения экспериментальных исследований был выбран набор реальной корпоративной переписки Enron[6], который широко применяют в статьях, посвящённых анализу текстовых данных, и конференциях по противодействию терроризму и компьютерной безопасности[7]. Для сравнения методов обнаружения аномалий использовалось значение площади под ROC-кривой — AUC (англ. *Area Under Curve*), которое является агрегированной характеристикой качества обнаружения/классификации, не зависящей от соотношения цен ошибок.

---

[6] Enron Email Dataset [Электронный ресурс]. — Электрон. дан. — [Б. м.] : 2015. — Режим доступа: http://www.cs.cmu.edu/~./enron/.

[7] Workshop on Link Analysis, Counterterrorism and Security [Электронный ресурс]. — Электрон. дан. — [Б. м.] : 2005. — Режим доступа: http://research.cs.queensu.ca/home/skill/proceedings/.

В **подразделе 3.2** описан предложенный подход прогнозирования тематической направленности пользователя. В данном подходе рассматривается пользовательский поток документов $x = \{(d, t)\}$, где документ d представляет объединённые текстовые данные пользователя, к которым он обращался за время $[t, t+\Delta t)$, при этом Δt выбирается достаточно «длительным». По потоку x строится тематическая модель поведения пользователя (L_m, W_k, H_k, T_n). Тогда поток x можно представить в виде множества упорядоченных пар $F = ((H^1, t_1), \ldots, (H^n, t_n))$, где $t_1 \le t_2 \le \ldots \le t_n$. Выборка F рассматривается как k-мерный временной ряд, по которому строится прогноз на следующие p шагов: $((H_f^{n+1}, t_{n+1}), \ldots, (H_f^{n+p}, t_{n+p}))$. После чего определяется решающая функция: $f((d, t_{n+j}), (L_m, W_k)) = \|H_f^{n+j} - h\|_1 = a_j$, где h — представление документа d в (L_m, W_k), $a_j \in \mathbb{R}$ — уровень аномальности обращения пользователя к контенту d за время $[t_{n+j}, t_{n+j}+\Delta t)$, $1 \le j \le p$. Здесь уровень аномальности соответствует отклонению тематической направленности пользователя от ожидаемых значений, поэтому чем меньше значение a_j, тем менее аномально поведение пользователя.

В **пункте 3.2.1** рассматриваются популярные методы прогнозирования временных рядов и разрабатывается собственный оригинальный метод на основе ортонормированной неотрицательной матричной факторизации. Идея предложенного метода состоит в нахождении взаимосвязей между элементами временного ряда, путём применения ортонормированной неотрицательной матричной факторизации к авторегрессионной матрице временного рядя порядка p. Матрица W_k будет описывать взаимосвязи среди p подряд идущих элементов временного ряда, тогда по уже известным $(p\text{-}1)$ значениям можно вычислить значение следующего элемента временного ряда.

Пункт 3.2.2 посвящен экспериментальному исследованию предложенного подхода к обнаружению аномального поведения пользователя на основе прогнозирования его тематической направленности. Из проведённой серии экспериментов, моделирующих задачу аутентификации пользователя на наборе электронных писем Enron, следует, что:

- предложенный подход показывает высокое качество обнаружения (полученные значения AUC находятся на уровне 0.9) аномального поведения даже при использовании стандартных методов прогнозирования;
- разработанный алгоритм прогнозирования временных рядов показал высокое качество прогнозирования по сравнению с другими популярными алгоритмами.

В **подразделе 3.3** приведено описание второго предложенного в работе подхода обнаружения аномального поведения пользователя на основе оценки принадлежности отдельных документов к характерным тематикам анализируемого пользователя. В данном подходе рассматривается поток документов $x = \{(d, t)\}$, где документ d соответствует тексту

документа, к которому пользователь обратился в момент времени t. По потоку x строится тематическая модель поведения пользователя (L_m, W_k, H_k, T_n). Тогда вычислять оценку степени принадлежности произвольного документа d' к тематикам пользователя было предложено как норму вектора h', являющегося представлением документа d' в (L_m, W_k), т.е. определяется решающая функция: $f(d', (L_m, W_k)) = \|h'\| = b$, при этом рассматривалось применение стандартных норм вектора: L^1, L^2, L^∞. Чем сильнее документ d' соответствует характерным тематикам анализируемого пользователя, тем менее аномально обращение данного пользователя к документу d'. Исходя из этого, оценку аномальности a можно вычислить как $g(b)$, где g — функция, задающая противоположную зависимость между величинами a и b. В работе использовалась функция $g(b) = -b$.

В **пункте 3.3.1** описывается процедура формирования экспериментального набора данных, состоящего из текстовых документов, прикреплённых к электронным письмам из набора Enron. **Пункты 3.3.2** и **3.3.3** посвящены экспериментальному исследованию рассмотренного подхода к обнаружению аномального поведения пользователя, при этом в **пункте 3.3.3** исследуется возможность применения предложенного метода удаления информационного шума из анализируемых документов. Из проведённой серии экспериментов, моделирующих задачу раннего обнаружения попыток хищения информации, следует, что:

– Предложенный подход обнаружения аномалий показывает высокое качество (полученные значения AUC находятся на уровне 0.9 и выше) выявления фактов работы пользователя с несвойственными для него документами. Предложенный подход показал лучшее качество обнаружения в сравнении с традиционными методами одноклассовой классификации.
– Предложенный метод удаления информационного шума из документов приводит к улучшению качества обнаружения аномалий (увеличение AUC от 2% до 6%) при этом объём обрабатываемых текстовых данных существенно сокращается (более чем на 70%). Улучшение результатов обнаружения наблюдается для методов классификации, использующих для векторного представления документов как пространство термов, так и пространство, формируемое современным методом doc2vec[8].

Четвёртая глава посвящена разработке архитектуры и программной реализации экспериментального образца программного комплекса (ЭО ПК) обнаружения аномального поведения пользователей при работе с текстовыми данными. Сведения об апробации ЭО ПК приведены в разделе «Общая характеристика работы» автореферата.

[8] Mikolov T. et al. Distributed representations of words and phrases and their compositionality // Advances in neural information processing systems. — 2013. — С. 3111-3119.

В **подразделе 4.1** проводится проектирование базовых сценариев функционирования ЭО ПК: сбор поведенческой информации (ПИ), построение индивидуальных поведенческих моделей, применение индивидуальных поведенческих моделей. Построение и применение пользовательских поведенческих моделей основаны на предложенных методах обнаружения аномального поведения пользователей.

В **подразделе 4.2** рассмотрена архитектура ЭО ПК, которая представляет собой мультиагентную систему, состоящую из следующих программных модулей (Рисунок 2):

1. *Агент мониторинга*. Программный агент, устанавливаемый на рабочее место пользователя, состоит из совокупности параллельно работающих модулей — *модуль сбора* и *модуль классификации*. *Модуль сбора* реализует сбор ПИ о работе пользователей с текстовыми данными. *Модуль классификации* служит для применения поведенческих моделей в режиме близком к реальному времени.
2. *Модуль консолидации поведенческой информации*. Программный агент, который консолидирует ПИ, получаемую от различных *агентов мониторинга*, в едином хранилище. В задачи модуля также входит предоставление доступа к единому хранилищу для формирования выборок ПИ, которые далее будут использоваться при создании и применении поведенческих моделей.
3. *Модуль построения индивидуальных поведенческих моделей*. На основе выборки ПИ модуль выполняет процедуру построения поведенческой модели, соответствующей одному из двух разработанных методов обнаружения аномального поведения пользователей. После формирования структур данных модели производится её сохранение в *хранилище моделей*.
4. *Модуль обнаружения аномального поведения* служит для применения поведенческих моделей к выборкам ПИ.
5. *Автоматизированное рабочее место (АРМ) аналитика*. Представляет графический интерфейс, реализующий базовые варианты использования ЭО ПК.

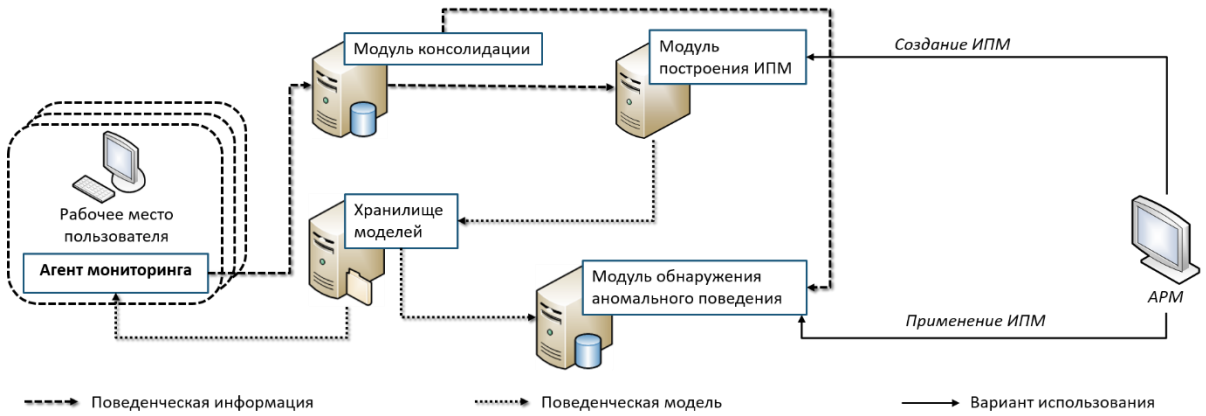

Рисунок 2 — Архитектура ЭО ПК.

В **подразделе 4.3** проведена комплексная оценка производительности разработанного ЭО ПК в части реализации базовых сценариев использования. Серия экспериментов с *агентом мониторинга* продемонстрировала, что работа агента существенно не сказывается на характеристиках наблюдаемого компьютера (в худшем случае загрузка ЦП 10%, ОЗУ 5.5 Мбайт). Оценочные эксперименты по построению и применению поведенческих моделей показали, что выполнение данных функций также не требует больших объёмом ОЗУ и не занимает длительное время.

В **заключении** сформулированы основные результаты диссертационной работы и перспективы дальнейшей разработки темы.

Основные результаты работы

1. Предложена новая модель представления потока текстовых документов в виде многомерного временного ряда, где каждая компонента ряда показывает изменение веса тематики во времени, при этом характерные тематики потока определяются с использованием методов ортонормированной неотрицательной матричной факторизации. Разработанная модель представления предназначена для решения задач анализа поведения пользователя при работе с текстовыми данными и фильтрации информационного шума из потоков текстовых документов.
2. Разработаны два новых алгоритма обнаружения аномального поведения пользователя при работе с текстовыми данными, использующих предложенное тематическое представление потока текстовых документов: алгоритм на основе анализа оценок принадлежности документов к характерным тематикам пользователя; алгоритм на основе анализа отклонений при прогнозировании тематических временных рядов пользователя.
3. Разработана архитектура и реализован экспериментальный образец мультиагентного программного комплекса, использующий предложенный комплекс алгоритмов для обнаружения аномального поведения пользователей по особенностям работы с текстовой информацией.

Публикации по теме диссертации

1. Петровский М.И., Глазкова В.В., Царёв Д.В. О выборе модели представления текстовой информации для задачи анализа и фильтрации Интернет-трафика // Математические методы распознавания образов: 13-я Всероссийская конференция. — М.: МАКС Пресс, 2007. — С. 519-522.

2. Машечкин И.В., Петровский М.И., Попов Д.С., Царёв Д.В. Латентно-семантический анализ в задаче автоматического аннотирования // Программирование. — Наука, 2011. — Т. 37. — № 6. — С. 67-77.

3. Tsarev D.V., Petrovskiy M.I., Mashechkin I.V., Popov D.S. Automatic text summarization using latent semantic analysis // Programming and Computer Software. — Springer, 2011. — Т. 37. — № 6. — С. 299-305.

4. Tsarev D.V., Petrovskiy M.I., Mashechkin I.V. Text Summarization Method Based on Normalized Non-Negative Matrix Factorization // 3rd International Conference on Mechanical and Electrical Technology (ICMET-China 2011). — ASME Press, 2011. — С. 563-568.

5. Tsarev D.V., Petrovskiy M.I., Mashechkin I.V. Using NMF-based text summarization to improve supervised and unsupervised classification // Hybrid Intelligent Systems (HIS), 2011 11th International Conference on. — IEEE, 2011. — С. 185-189.

6. Царёв Д.В. Исследование и разработка системы мониторинга потоков корпоративной электронной текстовой информации // Программные системы и инструменты. Тематический сборник №13. — М.: Изд-во факультета ВМиК МГУ, 2012. — С. 159-173.

7. Tsarev D.V., Petrovskiy M.I., Mashechkin I.V. Supervised and Unsupervised Text Classification via Generic Summarization // International Journal of Computer Information Systems and Industrial Management Applications. — 2013. — Т. 5. — С. 509-515.

8. Машечкин И. В., Петровский М. И., Царёв Д. В. Методы вычисления релевантности фрагментов текста на основе тематических моделей в задаче автоматического аннотирования // Вычислительные методы и программирование. — НИВЦ МГУ, 2013. — Т. 14. — № 1. — С. 91-102.

9. Герасимов С.В., Курынин Р.В., Петровский М.И., Попов И.С., Царёв Д.В., Шестимеров А.А. Технология оценки качества научно-технических документов // Программные системы и инструменты. Тематический сборник №14. — М.: Изд-во факультета ВМиК МГУ, 2013. — С. 158-171.

10. Герасимов С.В., Курынин Р.В., Машечкин И.В., Петровский М.И., Царёв Д.В., Шестимеров А.А. Инструментальные средства оценки качества научно-технических документов // Труды Института системного программирования РАН. — ИСП РАН, 2013. — Т. 24. — С. 359-380.

11. Tsarev D., Kurynin R., Petrovskiy M., Mashechkin I. Applying non-negative matrix factorization methods to discover user’s resource access patterns for computer security tasks // In Proceedings of the 2014 International Conference on Hybrid Intelligent Systems (HIS 2014). — New York, United States: IEEE Computer Society, 2014. — С. 43-48.

12. Машечкин И.В., Петровский М.И., Царёв Д.В. Применение методов интеллектуального анализа текстовой информации для предотвращения утечек данных // Программирование. — Наука, 2015. — № 1. — С. 32-43.

13. Tsarev D.V., Petrovskiy M.I., Mashechkin I.V., Popov D.S. Applying text mining methods for data loss prevention // Programming and Computer Software. — Springer, 2015. — Т. 41. — № 1. — С. 23-30.

14. Королев В.Ю., Корчагин А.Ю., Машечкин И.В., Петровский М.И., Царёв Д.В. Применение временных рядов в задаче фоновой идентификации пользователей на основе анализа их работы с текстовыми данными // Труды Института системного программирования РАН. — ИСП РАН, 2015. — Т. 27. — № 1. — С. 151-172.

15. Машечкин И.В., Петровский М.И., Царёв Д.В. Методы машинного обучения для анализа поведения пользователей при работе с текстовыми данными в задачах информационной безопасности // Вестник Московского университета. Серия 15: Вычислительная математика и кибернетика. — МГУ, 2016. — № 4. — С. 33-39.

16. Tsarev D.V., Petrovskii M.I., Mashechkin I.V. Machine Learning Methods for Analyzing User Behavior when Accessing Text Data in Information Security Problems // Moscow University Computational Mathematics and Cybernetics. — Springer, 2016. — Т. 40. — № 4. — С. 179-184.

17. Машечкин И. В., Петровский М. И., Поспелова И.И., Царёв Д. В. Методы автоматического аннотирования и выделения ключевых слов в задаче обнаружения экстремистской информации в сети Интернет // Современные информационные технологии и ИТ-образование. — 2016. — Т. 12. — № 1. — С. 188-200.

18. Mashechkin I., Petrovskiy M., Pospelova I., Tsarev D. Automatic summarization and keywords extraction methods for discovering extremist information on the internet // CEUR Workshop Proceedings (CEUR-WS.org): Selected Papers of the First International Scientific Conference Convergent Cognitive Information Technologies (Convergent 2016). — Т. 1763. — Moscow, Russia, 2016. — С. 188-198.

19. Интеллектуальная система оценки качества научно-технических документов [Текст] : пат. 132587 Рос. Федерация; дата рег. 20.09.2013.

20. Система мониторинга работы пользователей с информационными ресурсами корпоративной компьютерной сети на основе поведения пользователей [Текст] : свидетельство о гос. рег. ПО 2014616126 Рос. Федерация; дата рег. 11.06.2014.

21. Система мониторинга, теневого копирования и автоматического аннотирования текстовых данных при работе пользователя с электронными документами [Текст] : свидетельство о гос. рег. ПО 2016618914 Рос. Федерация; дата рег. 09.08.2016.