Методика проблемно-ориентированного анализа Big Data в режиме ограниченного времени

Д. В. Смирнов

***Аннотация* — Обсуждается методика организации выполняемого в режиме ограниченного времени поиска в Big Data вкраплений признаков вредоносных инсайдерских активностей. Методика апробирована в крупной индустриальной организации, операционная инфраструктура которой охватывает несколько тысяч серверов, сотни информационных ресурсов. В рамках реализации своих производственных функций этими информационными ресурсами постоянно пользуются несколько десятков тысяч сотрудников. Критически значимые ограничения, учет которых необходим при поиске признаков инсайдерских активностей, это динамически пополняемые операционные данные о характеристиках бизнес-активностей, данные мониторинга, информация об активностях эксплуатационного персонала и др. При этом динамически меняющимся объектом является также и профиль угроз, отражающий текущее состояние знаний о «природе» вредоносных инсайдерских активностей.**

В предложенной методике анализ данных ведется в режиме ограниченного времени, обеспечивая при этом изменяющиеся потребности текущей ситуации.

Представленная методика может быть обобщена для решения прикладных задач подобного типа. Работоспособность методики и разработанных для ее реализации программных средств демонстрируется на примере организации противодействия инсайдерским активностям крупного российского коммерческого банка.

***Ключевые слова* — Big Data, интеллектуальный анализ данных, ограниченное время, инсайдеры, информационная безопасность.**

I. Введение

Уже несколько десятилетий крупные предприятия используют различные информационные базы или хранилища, данные которых используются как для оперативного управления, так и для подготовки различных видов отчетности. Существенное снижение стоимости хранения одного терабайта данных сегодня – одна из основополагающих предпосылок к созданию единых хранилищ данных (Big Data). Централизация управления данными позволила значительно ускорить аналитическую и отчетную деятельность предприятий, способствовала созданию специальных управленческих моделей, а также их алгоритмических реализаций для встраивания в процессы управления. Вместе с тем Big Data стали областью концентрации всех данных предприятия и доступны аналитикам, работающих с данными, что привело к возрастанию рисков несанкционированного копирования или просмотра данных. Такой риск часто называют риском враждебного инсайдера.

В информационной системе для идентификации признаков вредоносных инсайдерских активностей имеются следующие сложности технического характера:

- профиль угроз (ПУ) является динамически изменяемым и содержит актуальные угрозы на текущий момент времени для объекта защиты;
- данные мониторинга являются динамически изменяемыми в процессе ведения бизнеса донными, из которых и «извлекаются» в соответствии с текущим вариантом ПУ признаки вредоносной инсайдерской активности;
- инфраструктура Big Data содержит тысячи серверов, сотни задействованных в операционной деятельности информационных ресурсов, десятки тысяч сотрудников эксплуатационного и обслуживающего персонала, что порождает проблемы сложности вычислений;
- формальный поиск вкраплений признаков инсайдерской деятельности в указанных условиях приводит к большому числу «ложных тревог» и большой вероятности пропуска вредоносной инсайдерской активности;
- информационная система должна оперировать в режиме ограниченного времени анализа данных и поддержки принятия управленческих решений (режим *not to be late*);
- информационная система должна поддерживать *гибкую реконфигурируемость* комплекса средств защиты, позволяющую обеспечить их соответствие изменившимся потребностям текущей ситуации;
- информационная система должна быть экономически эффективна, т.к. если система будет слишком затратной, то она неконкурентоспособна.

Таким образом, приходится иметь дело с задачей анализа больших объемов данных и поддержки принятия управленческих решений в рамках жестких ограничений по времени. Эффективное решение этих задач обеспечивается использованием совокупности проблемно-ориентированных средств анализа, объединенных на базе интеллектуального анализа данных (ИАД).

Статья получена 17 июля 2021 года.
Д. В. Смирнов, ПАО Сбербанк России, dvlsmirnov@sberbank.ru

II. ИТ-среда для реализации методики интеллектуального анализа данных и поддержки принятия решений для задач обсуждаемого типа

При формировании методики ИАД и поддержки принятия решений были учтены важнейшие особенности анализируемой ИТ-среды.

- На входе системы идентификации признаков вредоносной инсайдерской активности необходимо работать с гетерогенными «сырыми», не размеченными данными, в которых требуется организовать поиск информации, релевантной вредоносной активности инсайдеров.
- Необходимо опираться на постоянно накапливаемый опыт анализа, как успешных, так и ошибочных действий: постоянно обучаться и переобучаться в динамически меняющейся информационной среде.
- Обеспечивать эффективность действий персонала и технических систем по идентификации инсайдерских активностей как с учетом динамики изменения угроз, так и необходимости оперировать огромными объемами данных, а также необходимости рационально управлять ресурсами, поддерживать операционную и экономическую эффективность соответствующих систем информационной безопасности.

Обсуждаемая методика охватывает следующие базовые процедурные составляющие:

- *нормализация* «сырых» Big Data, ориентированная на использование инструментов *представления знаний*. В частности, использовать проблемно-ориентированные фреймы, формируемые с использованием данных из различных информационных пространств; ПУ состоит из типовых сценариев (ТС) инсайдерских активностей в различных информационных пространствах;
- ИАД используется как средство для реализации обучения на прецедентах в условиях эффекта появления новых данных (эффект *Open*), в том числе для выделения в них устойчивых зависимостей причинно-следственного типа;
- комбинирование условий применимости *статистических* и *дискретных* методов ИАД при идентификации и анализе аномалий в поведении объектов мониторинга: технических систем, персонала, и др.;
- *управление перебором* вариантов, позволяющих выделять признаки инсайдерских активностей;
- использование *проактивного* характера действий офицеров службы безопасности в процессе идентификации инсайдерских активностей и организации противодействия вредоносным воздействиям.

Обратимся к более детальному обсуждению элементов предлагаемой методики, а также разработанного для ее реализации программно-технического комплекса ИАД и процедуры принятия решений (ППР).

III. Способ анализа Big Data в режиме ограниченного времени

Характерные для Big Data эффекты *Big* и *Open* анализируемых данных изменили традиционное отношение к данным:

- если раньше данные помещались в различных базах, то с появлением технологий Big Data их стали хранить в централизованном хранилище (DWH);
- скорость пополнения DWH новыми загрузками данных при текущих объемах таких пополнений не позволяет вести централизованные реестры поисковых индексов для работы с данными разной структуры.
- традиционно [1], индексация документов для поисковых задач была в значительной мере ручным трудом. Сегодня актуальные размеры поисковых индексы увеличилась в миллионы раз и уже не могут формироваться в ручном режиме. Требуется автоматический подход к индексированию;
- для современных бизнес-задач поисковое индексирование должно происходить в ограниченное время, что в реальных бизнес-приложениях требует обработки десятков миллионов документов в сутки.

При разработке методики пристального внимания потребовала проблема структурирования исходных «сырых» данных, т.е. выделения в них данных, релевантных тем или иным поисковым потребностям, индексирования, организации поиска и др.

Эффекты Big Data усугубили процедурные сложности отбора релевантных сведений в первичных «сырых» данных, то есть проблему отбора в «сырой» неиндексированной информации релевантных тем или иным целям поиска данных и знаний.

Проблему поиска в первичных «сырых» данных предложено разделить на две подпроблемы: отбор релевантных данных для создания поисковых индексов и, собственно, сам поиск, который рассматривается как задача машинного обучения [2] по прецедентам, т.е. релевантных и нерелевантных ответов. Поисковая проблема уточняется как задача «реконструкции» частично-определенного примерами и контрпримерами отношения релевантности «цель поиска – ответы на поисковый запрос». Ошибки первого и второго рода используются для корректировки алгоритмов машинного обучения.

Дополнительно в методике появилась тема объяснения результатов. Сотрудники должны доверять предлагаемым им сведениям. Ошибки как первого, так и второго рода в части релевантности результатов целям поиска могут повлечь за собою тяжелые бизнес-последствия, отвечать за которые придется сотруднику, принявшему решение доверять этим результатам. Наличие неформального объяснения позволяет осознанно принять предлагаемые результаты поиска. Именно по этой причине в критичных приложениях требуется содержательное, использующее термины и понятия анализируемой предметной области объяснение решения, сформированного поисковым алгоритмом, а также удобный интерфейс для представления такого объяснения человеку-оператору.

Основная проблема создания методики состояла в организации пути поиска вкраплений признаков вредоносной инсайдерской активности в деятельности сотрудников крупного коммерческого банка (см. также [3]).

IV. Проблема обеспечения штатного режима функционирования крупного коммерческого банка в аспекте ИБ

В решаемой задаче программно-аппаратный комплекс хранения и обработки Big Data – это большой динамический объект размеров более 2000 серверов, расширяемый со скоростью пополнения 400+ серверов в год. В инфраструктуре этого комплекса Big Data работают более 6000 пользователей, и это число возрастает ежегодно примерно на 500 человек. К обсуждаемому комплексу Big Data подключены более 200 источников первичной информации (Oracle, MSSQL, Postgres базы, внешние базы партнеров и т.д.). В комплексе на текущий момент хранится более 40 петабайт данных. Этот объем данных возрастает примерно на 200 терабайт еженедельно.

Требуется фиксировать события, связанные с признаками инсайдеров, в ограниченное время, которое на текущий момент внутрикорпоративные регламенты определяют единицами секунд. Доля ложных срабатываний согласно действующим корпоративным регламентам должна быть менее 1/20. Иначе возникнет дефицит ресурсов реагирования на инциденты.

V. Реализация методики по идентификации признаков вредоносных инсайдерских активностей и организации противодействия им

Компьютерный инструмент ИА Big Data должен быть использован в качестве ассистента офицера безопасности при решении задач поиска признаков инсайдеров, организации и противодействия их вредоносной активности. Этот инструмент должен обеспечить постоянный сбор и анализ всех релевантных целям поиска признаков вредоносной инсайдерской активности сведений, как в материалах текущей работы службы безопасности, так и в других значимых источниках. Методика предлагает в качестве первого шага разработать и сопровождать в актуальном состоянии ПУ. Задача сопровождения ПУ в актуальном состоянии потребовала разработки специальной интеллектуальной подсистемы управления конфигурациями ПУ.

В качестве следующего шага методики необходимо создать механизмы фильтрации 300 Гб данных журналов, чтобы не обрабатывать нерелевантные данные и не переполнять поисковые индексы. Для этого потребовалось разработать алгоритмы и программные инструменты проблемно-ориентированного на поиск признаков вредоносной инсайдерской активности ИАД. Данная система-ассистент, сопровождая оперативного сотрудника через простой и понятный для него интерфейс, должна иметь все атрибуты “комфортного” поискового инструмента – нормализованные данные, возможности устранять опечатки и “ослышки”, представлять данные в понятной для человека форме и т.д.

VI. Жизненный цикл системы выявления признаков инсайдерской активности

Работа такой системы как объекта управления характеризуется присущим ей жизненным циклом. Целостное представление о компонентах жизненного цикла позволяет контролировать успех при решении задачи поиска признаков вредоносной инсайдерской активности в целом, помогая понять какие именно инструменты обработки данных применить и на каких этапах. Жизненный цикл позволяет выбрать управляющие воздействие в отношении конкретных данных в соответствующий момент времени.

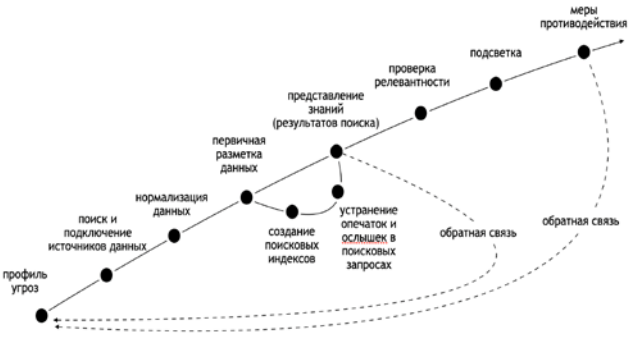

Рис.1. Жизненный цикл системы выявления признаков инсайдера

VII. Профиль угроз

В каждый момент времени актуальный ПУ представляет собою набор ТС. В исходном состоянии каждый ТС характеризуется текстовым описанием.

Текстовое описание преобразуется в специальный фреймовый вид. После выбора соответствующих средств формализованного представления знаний каждый ТС актуального ПУ преобразуется в удобный для компьютерной обработки вид. Например, в случае характеризации слотов таких фреймов признаками, принимающими булевские значения, каждый ТС может быть представлен как как битовая строка. Таким образом формируется матрица ***ПРИЗНАКИ*** x ***ТС***, столбцы которой (ТС) возможно сравнивать как битовые строки. Потребность оперировать с Big Data быстро и эффективно здесь отрабатывается за счет возможностей выполнять операции над ТС с помощью стандартных операций с битовыми строками, например, сравнение двух ТС за одну макрооперацию. Пример текстового описания приведен в работе [3].

Реализация методики отбора из имеющихся Big Data всех релевантных уже известным признакам доступа пользователей к защищаемым информационным ресурсам, демонстрирующих признаки вредоносной инсайдерской активности, потребовала разработки адекватных средств и специальных математических моделей анализа данных.

VIII. Использование профиля угроз в мониторинге пользователей

Процедуры поиска признаков инсайдерских активностей предполагают последовательное сравнение всех в текущий момент профилей доступов пользователей к защищаемым информационным ресурсам со всеми ТС текущего ПУ. В такой ситуации сокращение перебора

вариантов может быть обеспечено за счет «экономной» организации сравнений ТС с текущими профилями доступа пользователей. Для этого сперва выделяются все максимальные по вложению общие фрагменты описаний ТС, т.е. строится диаграмма сходств [3-6] описаний ТС, упорядоченных по взаимному вложению.

В теоретико-множественном случае это диаграмма взаимной вложимости множеств общих признаков подмножеств ТС из ПУ, далее каждый текущий профиль доступа сравнивается сперва с самыми общими сходствами, т.е. элементами нижней границы такой диаграммы. Затем, двигаясь по цепочкам частичного порядка этой диаграммы вверх к верхней границе, достигнуть необходимый ТС. Можно провести исчерпывающий анализ всех актуальных вариантов сравнений, исключая при этом повторное обращение к тем или иным сходствам, возникающее при сравнении по схеме каждый профиль доступов с каждым ТС.

Управление сравнениями с ТС в динамически изменяемой среде, как собственно Big Data, так и знаний о ПУ, – особая задача, которая потребовала создания специальной подсистемы, в которой аналитик может целенаправленно управлять приоритетами в анализе признаков и угроз. Так, например, продвижение по цепям частичного порядка диаграммы сходств ТС «снизу-вверх» позволяет не только сузить объем перебора релевантных конкретному отслеживаемому случаю фрагментов ТС, но и указать наиболее опасные варианты дальнейшего возможного развития ситуации во времени: «вверх» вдоль таких цепей частичного порядка. Это, в свою очередь, позволяет организовать управление необходимыми ресурсами, фокусируя в проактивном режиме внимание сотрудников службы безопасности на приоритетных угрозах.

Использование диаграммы сходств позволяет повысить скорость обработки данных. В теоретико-множественном случае, когда фрейм каждого ТС характеризуется множеством соответствующих признаков, диаграмма сходств строится с помощью операций пересечения множеств таких представляющих каждый ТС признаков. Т.е. в ней имеющиеся признаки ТС сгруппированы по совпадению, позволяющему в случае их вхождения в несколько ТС не проводить повторные их сравнения с текущим профилем доступов.

В предлагаемом процедурном подходе один раз строится диаграмма сходств и делаются проверки только с теми элементами, которые расположены на цепочках частичного порядка. В реальной практике, когда речь идет о миллиардах событий и о сотнях ТС, это позволяет получить значительный выигрыш в скорости принятия решений и объемах необходимых для этого вычислений по сравнению с методами прямого исчерпывающего перебора вариантов (*brute-force*). Дополнительный выигрыш в ресурсах и времени формируется за счет необходимости сопровождать динамические изменения «на ландшафте мониторинга»: постоянно расширяется и «поле» требующих анализа «сырых» данных, и сведения об актуальном спектре угроз. Методы сокращения перебора являются важнейшим этапом разработанной методики.

Созданный программно-технический инструментарий идентификации признаков вредоносной инсайдерской активности – это тоже программно-аппаратный комплекс Big Data, имеющий соответствующее аналитическое ядро и интерфейсы пользователей. В его архитектуре [3] задействованы и типовые программные инструменты: базы данных, серверы приложений, среды исполнения программного кода (Python, Java) и др. В аналитические ядро встроены алгоритмы фильтрации поступающих событий и обработка профиля угроз.

Реализующий методику программный комплекс имеет два различных интерфейса для поддержки проблемно-ориентированного представления знаний в первичном анализе исходных «сырых» данных, и вторичном представлении, которое служит для информационного сопровождения мониторинга сотрудниками службы безопасности.

При первичном поиске аналитик формализует знания из текущего ПУ, определяет необходимые данные (поля, идентификаторы и т.д.), а далее решает соответствующие задачи машинного обучения, где реконструирует требуемое отношение релевантности.

Вторичный поиск нужен для оперативного отображения релевантных запросов. Это своего рода «локальный Яндекс», где сотрудник может ввести ФИО, табельный номер или источник данных и посмотреть профиль сотрудника или подразделения с рассчитанными характеристиками. Например, использовать результаты скоринга различных характеристик аномалий, связанных с этим сотрудником в производственных процессах. Вторичный поиск – это обычный информационно-поисковый сервис.

IX. Эффективное сокращение данных для анализа

Особое внимание в методике нашла организация интерфейсов поиска. В интерфейсе первичного поиска аналитик отбирает то, что должно быть учтено при мониторинге угроз. То есть переводит в удобную для компьютерной обработки форму релевантные целям мониторинга знания об угрозах. В этот интерфейс в качестве специальных сервисов “удобным” образом подключены механизмы нормализации данных, где задействованы специально разработанные алгоритмы, ориентированные на оптимизацию переборов для соблюдения режима ограниченного времени.

Во вторичном поиске предлагается поисковый интерфейс с текстовым полем для ввода и кнопкой “искать”. Цель вторичного поиска – оперативно предоставлять информацию, включающую результаты работы алгоритмов машинного обучения, в простом и понятном виде для сотрудников, не имеющих продвинутых ИТ-навыков. Основная цель здесь – обеспечить ускорение (и упрощение) работы оперативных сотрудников. В созданный комплекс встроен ряд оригинальных алгоритмов. Например, в различных источниках поля и значения полей данных, как правило, называются по-разному. В частности, одно

и тоже наименование места или города в различных база данных может именоваться различными способами: так Москва может десятки различных написаний, в т.ч. - “MOSCOW”, “G.MOSKVA”, “MOSKVA”, “ГОРОД МОСКВА”, “МОСКВА, МОСКВОСКАЯ ОБЛАСТЬ” и т,д. Для сведения всех подобных «смысловых синонимов» к единому – «каноническому» представлению разработаны специальные алгоритмы нормализации представления данных. Выполняемая нормализация преследует три основные цели:

- уменьшить объемы данных, задействованных в реальном мониторинге,
- объединять (унифицировать разнородные) данные,
- представлять данные пользователю в удобной для него унифицированной форме.

Например, в одной из 30 разработанных витрин – ненормализованных названий доступов к данным – 167000 (уникальных – 3400), а нормализованных (официальных) названий доступов – уже 114000 (уникальных – 212). Нормализация в среднем снизила потребность в их хранении на 40% (114000/167000).

Также в представляемый комплекс «инструментов» ИАД встроены специальные, позволяющие варьировать детальность описания ТС, механизмы представления знаний об угрозах и средства статистического анализа аномалий в поведении объектов мониторинга [7-11]. В частности, инструменты выявления и статистического оценивания «значимости» случаев, когда сотрудники используют чужую учетную запись, или же их поведение отклоняется от поведения их коллег по ряду значимых признаков.

Дата	Кол-во событий в логах	1 ЭТАП ОБРАБОТКИ: Кол-во hdfs-audit.log логов (строк)	2 ЭТАП ОБРАБОТКИ: Кол-во логово после применения фильтров "src=/data" AND "cmd=getfileinfo" (elastic filter) (строк)	3 ЭТАП ОБРАБОТКИ: Кол-во логов после применения фильтра по тексту "part-" (elastic filter) (строк)	4 ЭТАП ОБРАБОТКИ: Кол-во событий в логах
12.04.2021	7 500 000 000	1 400 000 000	560 000 000	280 000 000	4 140 000
13.04.2021	7 500 000 000	1 400 000 000	560 000 000	280 000 000	644 390
14.04.2021	7 500 000 000	1 400 000 000	560 000 000	280 000 000	1 121 000
15.04.2021	7 500 000 000	1 450 000 000	580 000 000	290 000 000	1 370 000
16.04.2021	7 500 000 000	1 450 000 000	580 000 000	290 000 000	599 000
19.04.2021	7 600 000 000	1 450 000 000	580 000 000	290 000 000	6 410 000
20.04.2021	7 600 000 000	1 500 000 000	600 000 000	300 000 000	2 460 000
21.04.2021	7 600 000 000	1 500 000 000	600 000 000	300 000 000	1 340 000
22.04.2021	7 600 000 000	1 500 000 000	600 000 000	300 000 000	1 470 000
23.04.2021	7 600 000 000	1 500 000 000	600 000 000	300 000 000	1 280 000
24.04.2021	7 600 000 000	1 500 000 000	600 000 000	300 000 000	803 300
25.04.2021	7 600 000 000	1 500 000 000	600 000 000	300 000 000	2 360 000
26.04.2021	7 600 000 000	1 500 000 000	600 000 000	300 000 000	1 950 000
27.04.2021	7 600 000 000	1 500 000 000	600 000 000	300 000 000	1 540 000
28.04.2021	7 600 000 000	1 500 000 000	600 000 000	300 000 000	1 530 000
29.04.2021	7 600 000 000	1 500 000 000	600 000 000	300 000 000	882 800
30.04.2021	7 600 000 000	1 500 000 000	600 000 000	300 000 000	1 280 000
10.05.2021	7 700 000 000	1 600 000 000	640 000 000	320 000 000	6 610 000
11.05.2021	7 700 000 000	1 600 000 000	640 000 000	320 000 000	2 840 000
12.05.2021	7 700 000 000	1 600 000 000	640 000 000	320 000 000	423 300

Таб.1. Эффект алгоритма фильтрации данных (отфильтрованы избыточные данные)

Если брать в качестве инструмента вторичного поиска известные поисковые дистрибутивы, например, такие как Elastic Search или Sphinx, то по нашему опыту они перестают функционировать в штатном режиме на при размерах индекса более 50 млн. записей. Отсюда возникла потребность в разработке алгоритма фильтрации данных, который состоит из 4 этапов и гарантирует, что ни одна релевантная запись из лог файла не будет отфильтрована.

В результате работы алгоритма фильтрации логов получается, что на каждом этапе фильтрации количество строк в лог-файлах уменьшается на порядок (Таб. 1), тем самым уменьшается объем поиска. Причем алгоритм фильтрации не отбрасывает ни одной полезной строки.

В результате данной работы был реализован анализ нескольких ПУ. Например, в одном из реализованных ПУ требовалось выявлять технологические учетные записи (TUZ), под которыми работают не процессы, а сотрудники (PUZ), тем самым скрывают свои следы обращения к данным (сценарий поведения - “*маскировщики*”, см. Рис. 2).

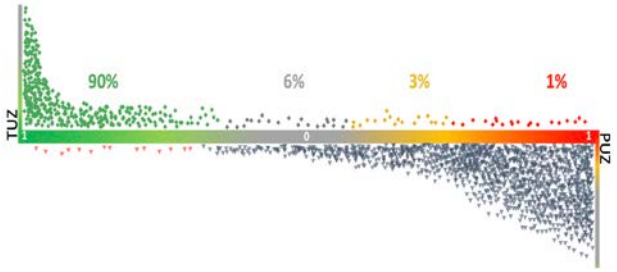

Рис.2. Результаты работы ПУ при выявлении технологических учетных записей, которые используют сотрудники

Для создания данного ПУ были использованы 1316 признаков, характеризующую учетную запись. В результате анализа ПУ были обнаружены 102 технологические учетные записи (TUZ), каждая из которых использовалась некоторыми сотрудниками для доступа к данным, а не к процессам.

На представленные выше программные инструменты интеллектуального анализа данных оформлены государственные свидетельства о Регистрации программы для ЭВМ и базы данных.

X. Заключение

Основное содержание методики может быть представлено следующим образом.

- Методика идентификации признаков вредоносных инсайдерских активностей и противодействия их влиянию.
- Методика сокращения сложности поисковых алгоритмов.
- Методика нормализации данных.

Для апробации методики создан программный комплекс, реализующий разработанную методику.

Библиография

[1] Ф.У. Ланкастер, *Информационно-поисковые системы. Характеристики, испытание и оценка*, Москва: Мир, 1972. 307 с.

[2] M. Kubat, *An Introduction to Machine Learning*, Springer, 2017. 348 p.

[3] Д. В. Смирнов, А. А. Грушо, М. И. Забежайло, Е. Е. Тимонина, “Система сбора и анализа информации из различных источников в условиях Big Data,” *International Journal of Open Information Technologies*, Т. 9, № 4, С. 64-74, 2021.

[4] М. И. Забежайло, “О некоторых возможностях управления перебором в ДСМ-методе,” *Искусственный интеллект и принятие решений*, Часть I: № 1, С. 95-110, Часть II: № 3, С. 3- 21, 2014.

[5] М. И. Забежайло, “О некоторых оценках сложности вычислений в ДСМ-рассуждениях,” Искусственный интеллект и принятие решений, Часть I: №1, С. 3-17, Часть II: №2. С. 3-17, 2015.

[6] А. А. Грушо, М. И. Забежайло, А. А. Зацаринный, Е. Е. Тимонина, “О некоторых возможностях управления ресурсами при организации проактивного противодействия компьютерным атакам,” Информатика и ее применения, Т. 12, № 1, С. 62-70, 2018.

[7] А. А. Грушо, М. И. Забежайло, Д. В. Смирнов, Е. Е. Тимонина, "О комплексной аутентификации," Системы и средства информ., Т. 27, Вып. 3, С.4–11, 2017.
[8] А. А. Грушо, М. И. Забежайло, Д. В. Смирнов, Е. Е. Тимонина, "Модель множества информационных пространств в задаче поиска инсайдера," Информатика и ее применения, Т. 11, № 4, С. 65-69, 2017.
[9] А. А. Грушо, Н. А. Грушо, М. И. Забежайло, Д. В. Смирнов, Е.Е. Тимонина, "Параметризация в прикладных задачах поиска эмпирических причин," Информатика и ее применения, Т. 12, № 3, С. 62-66, 2018.
[10] А. А. Грушо, М. И. Забежайло, Д. В. Смирнов, Е. Е. Тимонина, С. Я. Шоргин," Методы математической статистики в задаче поиска инсайдера," Информатика и ее применения, Т. 14, Вып. 3, С. 71-75, 2020.
[11] А. А. Грушо, М. И. Забежайло, Д. В. Смирнов, Е. Е. Тимонина, "О вероятностных оценках достоверности эмпирических выводов," Информатика и ее применения, Т. 14, Вып. 4, С. 3-8, 2020.

Methodology of Problem-Oriented Big Data Analysis in Limited Time Mode

D. V. Smirnov

***Abstract* — The methodology of organization of search in Big Data, performed in the mode of limited time, of signs of malicious insider activities is discussed. The methodology is tested in a large industrial organization, the operating infrastructure of which covers several thousand servers, hundreds of information resources. As part of their operational functions, several tens of thousands of employees are constantly using these information resources. Critical limitations, which must be taken into account when looking for insider activity characteristics, are dynamically replenished operational data on business activity characteristics, monitoring data, information on operational personnel activities, etc. At the same time, a dynamically changing object is also a threat profile, reflecting the current state of knowledge about the "nature" of malicious insider activities.**

In the proposed methodology, the analysis of data is carried out in the mode of limited time, while ensuring the changing needs of the current situation.

The presented technique can be generalized to solve tasks of this type. The operability of the methodology and the software developed for its implementation is demonstrated by the example of the organization of counteracting malicious insider activities in large Russian commercial bank.

***Keywords* — Big Data, intellectual data analysis, limited time, insiders, information security.**

REFERENCES

[1] F.W. Lancaster, *Information retrieval systems; characteristics, testing, and evaluation*. New York, Wiley, 1968.

[2] M. Kubat, *An Introduction to Machine Learning*. Springer, 2017. 348 p.

[3] D. V. Smirnov, A. A. Grusho, M. I. Zabezhailo, E. E. Timonina, "System for collecting and analyzing information from various sources in Big Data conditions," *International Journal of Open Information Technologies*, vol. 9, no. 4, pp. 64-74, 2021. Available: http://injoit.org/index.php/j1/article/view/1099

[4] M. I. Zabezhailo, "To some new possibilities to control computational complexity of hypotheses," *Scientific and Technical Information Processing*, Part I: no. 1, pp. 95-110, Part II: no. 3, pp. 3- 21, 2014.

[5] M. I. Zabezhailo, "To the computational complexity of hypotheses generation in JSM-method," *Scientific and Technical Information Processing*, Part I: no. 1, С. 3-17, Часть II: no. 2. С. 3-17, 2015.

[6] A. A. Grusho, M. I. Zabezhailo, A. A. Zatsarinny, E. E. Timonina, "On some possibilities of resource management for organizing active counteraction to computer attacks," *Informatics and Applications*, vol. 12, no. 1, pp. 62-70, 2018.

[7] A. A. Grusho, N. A. Grusho, M. I. Zabezhailo, D. V. Smirnov, E. E. Timonina, "About complex authentication," *Systems and Means of Informatics*, vol. 27, no. 3, pp. 3-10, 2017.

[8] A. A. Grusho, M. I. Zabezhailo, D. V. Smirnov, E. E. Timonina, "The model of the set of information spaces in the problem of insider detection," *Informatics and Applications*, vol. 11, no. 4, pp. 65-69, 2017.

[9] A. A. Grusho, N. A. Grusho, M. I. Zabezhailo, D. V. Smirnov, E. E. Timonina, "Parametrization in Applied Problems of Search of the Empirical Reasons," *Informatics and Applications*, vol. 12, no. 3, pp. 62-66, 2018.

[10] A. A. Grusho, M. I. Zabezhailo, D. V. Smirnov, E. E. Timonina, S. Ya. Shorgin, "Mathematical statistics in the task of identifying hostile insiders," *Informatics and Applications*, vol. 14, no. 3, pp. 71-75, 2020.

[11] A. A. Grusho, M. I. Zabezhailo, D. V. Smirnov, E. E. Timonina, "On probabilistic estimates of the validity of empirical conclusions," Informatics and Applications, vol. 14, no. 4, pp. 3-8, 2020.