Обработка больших данных в системах мониторинга банковских транзакций

М.Ю. Сапожникова
Факультет информатики и робототехники
Уфимский государственный авиационный технический университет
Уфа, Россия
e-mail: sapozhnikova.maria.pro@yandex.ru

А.М. Вульфин
Факультет информатики и робототехники
Уфимский государственный авиационный технический университет
Уфа, Россия
e-mail: vulfin.alexey@gmail.com

М.М. Гаянова
Факультет информатики и робототехники
Уфимский государственный авиационный технический университет
Уфа, Россия

А.В. Чуйков
Факультет информатики и робототехники
Уфимский государственный авиационный технический университет
Уфа, Россия

Д.В. Курамшин
Факультет информатики и робототехники
Уфимский государственный авиационный технический университет
Уфа, Россия
autorbdja@mail.ru

Аннотация[1]

Для повышения эффективности обнаружения мошеннических банковских транзакций предлагается структура системы анализа данных пользовательской среды с целью выявления мошеннических действий. Система сбора и анализа информации о пользовательской среде позволяет накапливать данные, размечать прецеденты в ручном и автоматическом режимах и создавать базу данных примеров для обучения классификаторов. Для применения инструментов интеллектуального анализа данных необходимо реализовать интерфейс сбора данных, хранения и доступа к информации. Для работы со значительным объемом накопленных данных требуется использование специальных инструментов (фреймворков и аппаратных платформ) для обработки больших данных. В статье представлен анализ существующих программных и аппаратных средств для распределенной обработки слабо структурированных данных банковских транзакций (фреймворки: Hadoop, Apache Spark). Разработаны структура и рекомендации по развертыванию программно-аппаратного стенда для тестирования алгоритмов обнаружения финансового мошенничества на основе методов интеллектуального анализа данных в составе распределенной системы обработки банковских.

Труды Шестой всероссийской научной конференции "Информационные технологии интеллектуальной поддержки принятия решений", 28-31 мая, Уфа-Ставрополь, Россия, 2018

1. Введение

Проникновение информационных технологий во все сферы человеческой жизни создает основу формирования новых условий функционирования рынка. В этих условиях стало возможным развитие так называемой цифровой экономики. Ключевыми факторами экономической деятельности становятся электронные технологии и услуги и представление в цифровом виде многоотраслевых данных больших объемов [1, 2]. Электронная торговля составляет значимый институт в этой отрасли экономики, проникает во все большее количество правоотношений, складывающихся в сфере торговле в электронной форме. Наблюдается бурный рост сферы финансовых технологий: внедрение технологий искусственного интеллекта, машинного обучения, анализа больших данных для повышения эффективности взаимодействия всех участников правоотношений [3, 4, 5].

Важным аспектом функционирования цифровой экономики является обеспечение информационной

и экономической безопасности бизнеса, защиты персональных данных. Следствием стремительного развития финансовых технологий во всем мире стало возросшее количество мошеннических действий, совершаемых в электронной среде. По данным Центрального банка РФ на 2014 год доля мошеннических операций в интернет-банкинге составила 63 %, а за последние 2 года – выросла в 5,5 раз и составила 93 % всех преступлений, связанных с хищением средств со счетов держателей карт [6, 7].

На сегодняшний день применение технологий обработки больших данных и методов интеллектуального анализа данных является важным элементом системы противодействия финансовому мошенничеству. Например, внедрение Big Data позволило HSBC повысила эффективность распознавание мошеннических инцидентов в 10 раз [8]. Антифрод система VISA помогает предотвратить мошеннические платежи на сумму 2 млрд долларов США ежегодно [9].

2. Анализ данных пользовательского окружения в составе системы обнаружения мошеннических транзакций

Система мониторинга транзакций (СМТ) или антифрод-система (от английского anti-fraud, fraud - мошенничество) – специализированный программный или программно-аппаратный комплекс, обеспечивающий мониторинг, обнаружение мошеннических действий, а также обеспечивающий поддержку принятия решения по обнаруженной незаконной операции

Наиболее перспективным решением на сегодняшний день является применение технологий определения пользовательского окружения в сочетании с методами машинного обучения. Применение машинного обучения – необходимый критерий, поскольку собирается достаточно большой объем информация о пользовательском окружении и применение правил к этим данным становится невозможным. Как обсуждалось ранее, классические методы обнаружения мошеннических действий не могут достаточно точно ответить на вопрос: действительно ли данное действие совершил пользователь? Существует большое число способов получения незаконного доступа к аккаунту пользователя: фишинг, вишинг, фарминг, мобильное мошенничество, а также другие способы, связанные с методами социальной инженерии [10, 11, 12]. Для анализа больших объемов собираемых о пользовательском окружении данных целесообразно использовать подходы на основе интеллектуального анализа данных. Применение алгоритмов ИАД при решении задачи распознавания мошеннических транзакций и анализа данных пользовательского окружения представлено здесь [11, 12]. Наибольшая эффективность достигается с использованием комбинации различных алгоритмов (stacking-bagging) и применением технологий больших данных (Hadoop Processing) [11]. Это объясняется тем, что в исходном виде эти алгоритмы уже не способны решить существующие задачи в виду возрастающих объемов обрабатываемых данных. Возникает потребность модифицировать эти алгоритмы, комбинировать их для получения приемлемого результата, а также применять технологии, способные обрабатывать огромные объемы накапливаемой информации.

Целью данного исследования является разработка инфраструктуры для сбора и анализа данных пользовательского окружения в составе системы обнаружения мошеннических транзакций на основе технологий обработки больших данных

Для достижения поставленной цели были сформулированы следующие задачи:

- Разработка структуры системы для сбора и анализа данных пользовательского окружения на основе технологий обработки больших данных.
- Разработка структурной и функциональной схемы обработки данных пользовательского окружения для тестирования алгоритмов выявления финансового мошенничества на основе ИАД в составе системы распределенной обработки данных банковских транзакций.

3. Структура системы сбора и анализа информации о пользовательском окружении

Технологии дистанционного банковского обслуживания для доступа к счетам и операциям через веб-браузер не требуют установки клиентской части ПО и получили очень широкое распространение [13, 14]. Пользователь совершает определенные манипуляции в веб-браузере, взаимодействующим с интерфейсом Frontend-сервера. Frontend-сервер формирует набор данных о пользовательском окружении и передает данные о действиях пользователя по инициализации транзакции на Backend-сервер системы дистанционного банковского обслуживания (ДБО) и затем в автоматизированную банковскую систему (АБС) банка для проведения расчетов [13, 14, 15]. Backend-сервер передает данные транзакции и собранные данные о пользовательском окружении для анализа в антифрод-систему. В случае признания легитимности транзакции данные передаются на сервера АБС, в противном случае Backend-сервер отказывает пользователю в совершении операции. Антифрод-система оценивает риск текущей транзакции и в случае превышения некоторого порогового значения запускает дополнительные механизмы проверки легитимности транзакции [16]:

- автоматизированные способы дополнительной аутентификации транзакции
    - sms / push-уведомление;

    - просьба ответить на контрольные вопросы;
- ручные способы дополнительной аутентификации транзакции
    - телефонный звонок специалиста службы безопасности пользователю.

В данной архитектуре модуль управления пользовательской сессией (МУПС) – основной элемент системы обнаружения мошеннических транзакций (антифрод системы). Модуль выполняет анализ данных транзакции и данных пользовательского окружения (ДПО), собираемых скриптом на стороне клиента. Внедряемый скрипт формирует набор данных о пользовательском окружении [17]. Обобщенная структура собираемых ДПО следующая:

- Color depth
- Document size
- Screen size
- Time zone offset
- Fonts
- Plugins
- IP-address
- Number of processor cores
- UserAgent.

Основная задача модуля – классифицировать текущую сессию и реализуемую в ее рамках транзакцию (легитимная транзакция или мошеннические действия) на основе композиции методов анализа: сигнатурного и автоматического. Если формируемая модулем оценка легитимности транзакции ниже порогового значения, задействуется дополнительный механизм аутентификации пользователя.

Модуль сигнатурного анализа позволяет использовать экспертные знания и их формализацию в виде системы продукционных правил вида «ЕСЛИ-ТО». Основная задача данного модуля – классифицировать данные пользовательского окружения и/или имеющиеся данные о банковской транзакции с целью выявления мошеннических действий. Особенностью модуля является использование унифицированной сигнатурной базы на основе системы продукционных правил, что позволяет встроить в систему механизм объяснения принимаемого решения в понятной для эксперта по безопасности форме. Первоначально база модуля сигнатурного анализа содержит типовые шаблоны мошеннических и легитимных транзакций и ДПО в соответствующих случаях. Пополнение базы сигнатур возможно, как в ручном режиме через интерфейс модуля ручного анализа, так и с помощью выявляемых модулем автоматического анализа сигнатур в режиме обработки новых данных. Анализ накопленных данных под контролем аналитика позволяет выявлять новые продукционные правила в автоматическом режиме на основе технологий ИАД.

Задачей модуля автоматического анализа является автоматическая классификация данных на основе методов ИАД, например, с помощью нейросетевого классификатора, обученного на промаркированной экспертом базе совершенных транзакций [11]. На этапе подготовки данных для применения алгоритмов ИАД необходимо решить следующие задачи:

- features selection – отбор наиболее значимых признаков для принятия классификационного решения.
- features transformation – заполнение пропусков в данных, удаление выбросов и фильтрация шумовых компонент.
- features extraction – преобразование отобранных признаков в новое пространство признаков для подачи на вход классификатора;

В предыдущих работах авторов [11, 12] исходное пространство признаков включало 40 параметров. После оценки имеющихся значений параметров и их распределения из исходных данных были удалены 12 параметров. В результате экспертного анализа результатов преобразования признакового пространство были добавлены новые нелинейные признаки – комбинации исходных, характеризующие возможные сочетания некоторых имеющихся параметров.

Ядром модуля автоматического анализа является нейронная сеть – «черный ящик» – позволяющая отнести текущий вектор признаков, выделенных из собираемых о пользовательском окружении данных, к одному из ранее определенных классов. Предложено деление на следующие классы:

- Пользовательская система находится под удаленным управлением
- Пользовательская система использует механизмы анонимизации действий
- Пользовательская система не содержит подозрительных элементов

Результаты работы модулей сигнатурного и автоматического анализа сопоставляются. Если вердикты не совпадают, то может быть привлечен эксперт антифрод-системы для ручного анализа прецедента. По мере накопления данных о совершаемых транзакциях и ДПО происходит доообучение нейросетевого классификатора. По мере формирования группы прецедентов, параметры которых не укладываются в текущую схему сигнатурных правил «ЕСЛИ-ТО», извлекаются новые правила, пополняющие существующую базу сигнатур. Если же экспертом добавляются новые сигнатуры, то происходит анализ текущей базы промаркированных прецедентов с целью обновления меток классов и переобучения нейросетевого классификатора.

Модуль ручного анализа предназначен для корректировки разметки имеющейся базы данных о пользовательском окружении и транзакциях с целью формирования обучающей выборки для нейросетевого классификатора. Модуль позволяет эксперту анализировать решение системы по каждому из прецедентов и корректировать его в случае, ошибочных срабатываний.

Модули управления АФС позволяет отслеживать основные показатели работы АФС, анализировать журнал работы системы и отлаживать взаимодействие сигнатурного и автоматического модулей.

Таким образом, представлена структура системы сбора и обработки данных пользовательского окружения в составе АФС. Ключевым элементом системы является модуль интеллектуального анализа данных. Алгоритмы анализа должны быть применимы в условиях «больших данных» (совокупности подходов, инструментов и методов обработки структурированных и неструктурированных данных огромных объёмов и значительного многообразия для получения воспринимаемых человеком результатов, эффективных в условиях непрерывного прироста, распределения по многочисленным узлам вычислительной сети) [18].

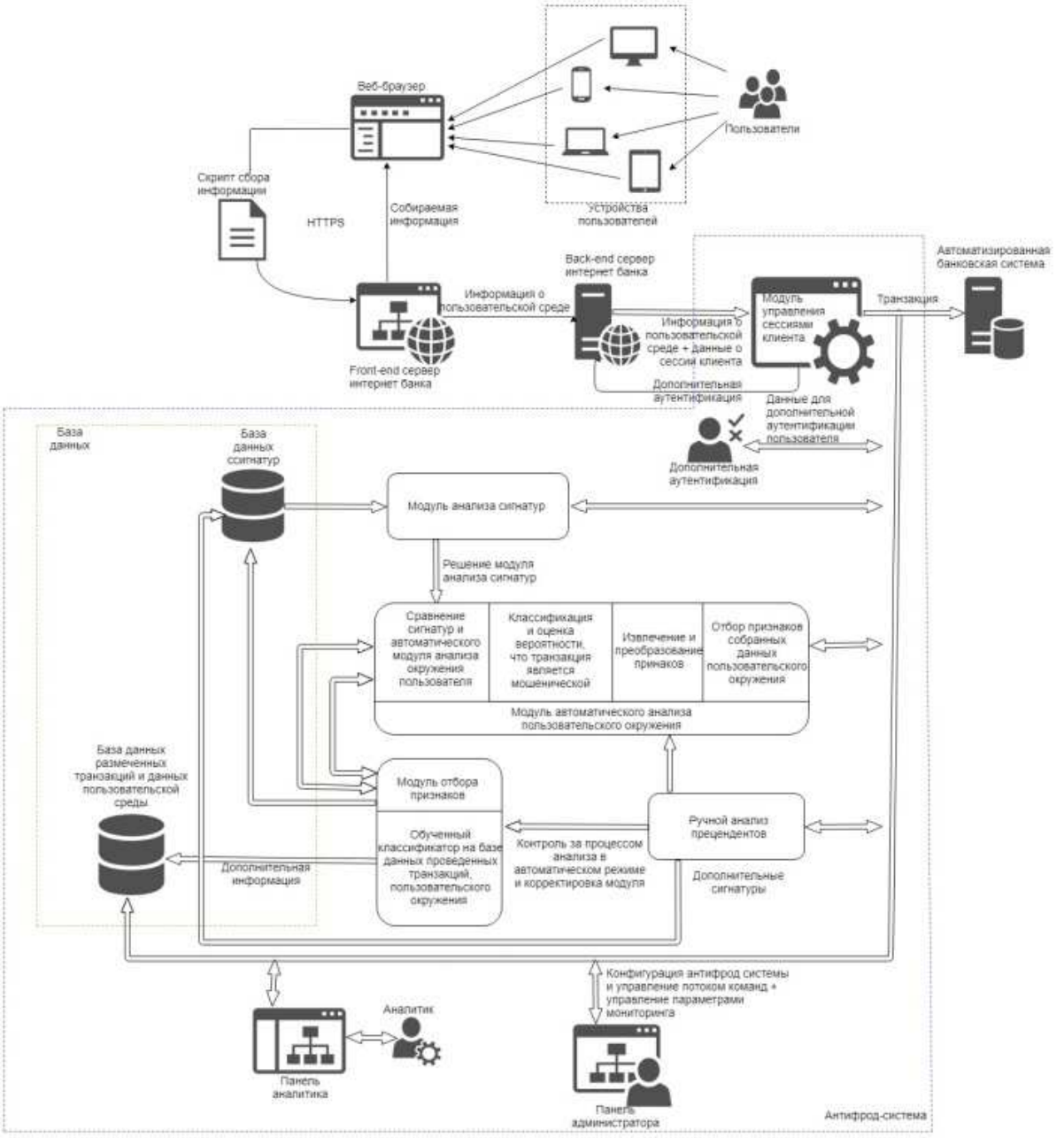

Рис. 1. Структура системы анализа ДПО и транзакции в составе АФС

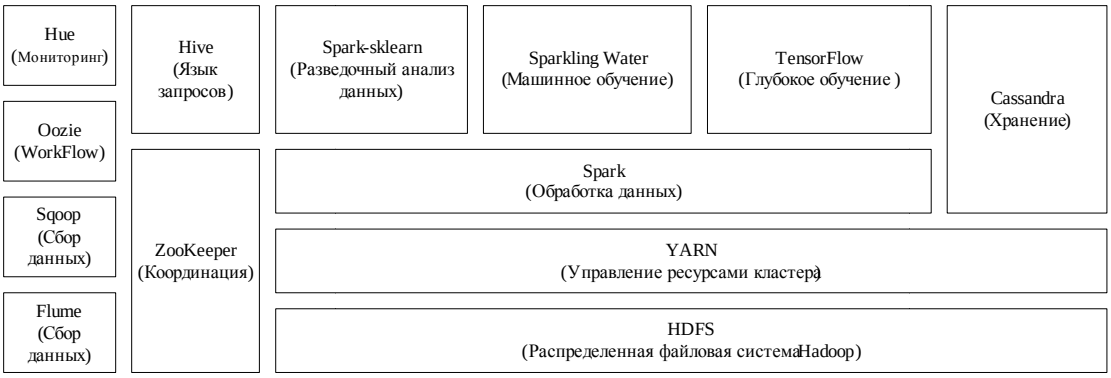

Рис. 2. Структура кластера Hadoop

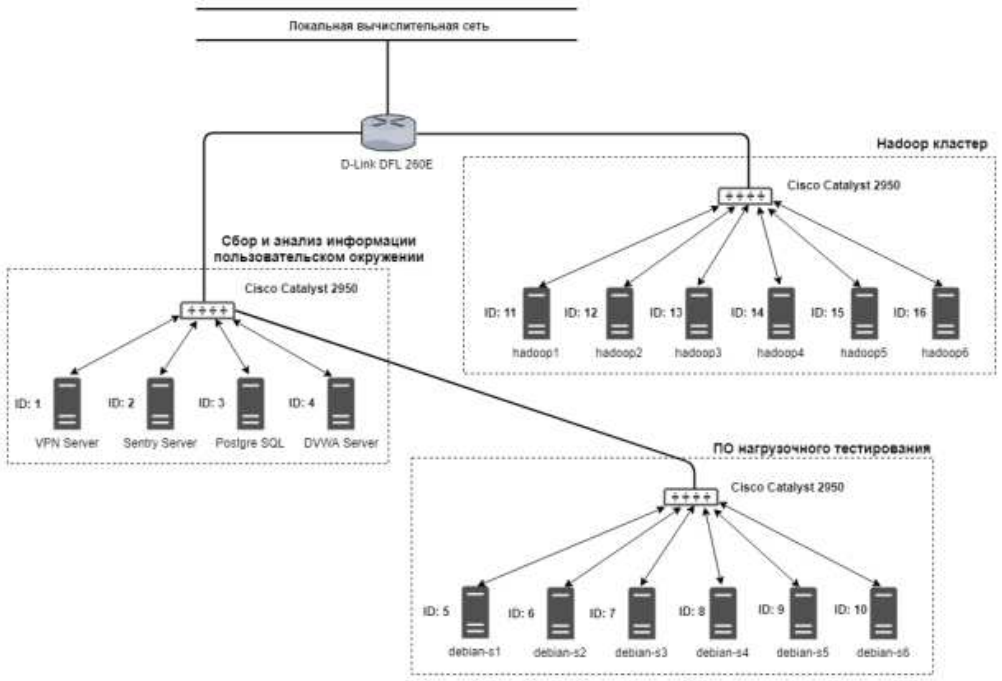

Рис. 3. Структура программно-аппаратного стенда для тестирования алгоритмов обнаружения финансового мошенничества

4. Проектирование структурной и функциональной схемы обработки «больших данных» пользовательского окружения в составе системы распределенной обработки данных банковских транзакций

Реализация алгоритмов выявления финансового мошенничества на основе ИАД банковских транзакций в составе системы распределенной обработки данных банковских транзакций требует решения ряда задач, связанных с проектирование и развертыванием соответствующей инфраструктуры для хранения и обработки накапливаемых данных.

На сегодняшний день существует множество инструментов распределенной обработки данных банковских транзакций (фреймворки: Hadoop, Apache Spark, ClickHouse, ElasticSearch, Splunk Free) [19, 20, 21, 22, 23]. Предлагаемая структура системы распределенной обработки данных банковских транзакций представлена на рис. 2.

Основным элементом системы распределенной обработки данных банковских транзакций является распределенная файловая система. Наиболее популярной на сегодняшний день является HDFS [19].

Следующий элемент системы обработки больших данных – инфраструктура распределенного программирования и машинного обучения. Ядром этого элемента является Spark – инфраструктура кластерных вычислений, сходная в MapReduce [19]. В состав данной инфраструктуры входит и инструмент машинного обучения MLLib, позволяющий реализовать инструменты ИАД накапливаемых данных.

Таблица 1
Приложения в экосистеме Hadoop

Назначение	Название	Краткое описание
Хранение	HDFS	Распределенная файловая система
	Cassandra	Система управления базой данных NoSql
Управление ресурсами кластера	YARN	Операционная система для big data приложений
Обработка данных	Spark	Ядро для обработки big data
Машинное обучение	Spark-sklearn	Библиотека Scikit-learn library интегрированная с Apache Spark для разведочного анализа данных
	Sparkling Water	Библиотека H2O интегрированная Apache Spark для машинного обучения в Hadoop
	TensorFlow	Библиотека TensorFlow интегрированная в Apache Spark для глубокого обучения в Hadoop
Координация задач	Zookeeper	Приложение для оддержания конфигурации, наименования кластеров и пр.
Доступ к данным	Hive	Приложения для доступа кданным с помощью SQL-подобных запросов
Сбор данных	Sqoop	Приложение для передачи данных между реляционными БД в Hadoop
	Flume	Приложение для обмена неструктурированным и данными в Hadoop
Управление потоком операций	Oozie	Приложение для управления потом заданий различным приложений в экосистеме Hadoop
Мониторинг	Hue	Web интерфейс для мониторингаHadoop

Для непосредственного хранения накапливаемых данных предлагается использовать решения из семейства New-SQL [24]. Детальное описание вспомогательных элементов и их функций в составе системы распределенной обработки данных банковских транзакций представлена в табл. 1. Структура программно-аппаратного стенда для тестирования алгоритмов выявления финансового мошенничества на основе ИАД в составе системы распределенной обработки данных банковских транзакций на основе стека технологий обработки Big Data представлена на рис. 3.

Модуль сбора и анализа данных представляет собой: комплекс ПО Sentry [34] для сбора логов скрипта клиентской стороны, сервис Gitlab [35] для организации совместной работы над исходным кодом реализуемых алгоритмов анализа, DVWA (Damn Vulnerable Web Application) для тестирования скрипта сбора данных по пользовательском окружении.

Модуль нагрузочного тестирования предназначен для автоматизации сбора базы данных о пользовательском окружении. Для распределенной обработки данных банковских транзакций предназначен Hadoop-кластер, на котором развернуто программное обеспечение из табл. 1. Типовая конфигурация используемого серверного парка машин приведена в табл. 2.

Таблица 2
Параметры серверов

ID	Сектор	Конфигурация	ОС
1	Сбор и анализ информации о пользовательском окружении	2x3.4 GHz / 4GB	debian 8.2
2		2x3.4 GHz / 2GB	debian 8.2
3		2x3.0 GHz / 1GB D	ubuntu 16.04
4		1x2.4 GHz / 1GB	ubuntu 16.04
5	Модуль нагрузочного тестирования	2x3.0 GHz / 3GB	debian 9.2
6		2x3.0 GHz / 3GB	debian 9.2
7		2x3.0 GHz / 4GB	debian 9.2
8		2x3.0 GHz / 3GB	debian 9.2
9		2x3.0 GHz / 4GB	debian 9.2
10		2x3.0 GHz / 3GB	debian 9.2
11	Hadoop кластер	2x3.4 GHz / 12GB	ubuntu 14.04
12		2x3.4 GHz / 12GB	ubuntu 14.04
13		2x3.4 GHz / 8GB	ubuntu 14.04
14		2x3.4 GHz / 6GB	ubuntu 14.04
15		2x3.2 GHz / 6GB	ubuntu 14.04
16		2x3.0 GHz / 6GB	ubuntu 14.04

5. Заключение

Основной проблемой повышения эффективности СМТ является недостаточный объем фиксируемых параметров, передаваемых с клиентской стороны онлайн-банкинга в процессинговый центр, и несовершенство методов и алгоритмов сигнатурного анализа в силу низких возможностей по адаптации и гибкой настройке.

Наиболее перспективным решением на сегодняшний день является применение технологий определения пользовательского окружения в сочетании с методами машинного обучения в составе СМТ. Применение машинного обучение – необходимый критерий, поскольку собирается большой объем информация о пользовательском окружении и применение правил к этим данным становится затруднительным. Алгоритмы анализа должны быть применимы в условиях «больших данных». В работе предложена инфраструктура для сбора и анализа данных пользовательского окружения в составе системы обнаружения мошеннических транзакций на основе технологий обработки больших данных.

Благодарности

Данное исследование выполнено при поддержке гранта РФФИ № 17-48-020095.

Список используемых источников

1. Teoh C.S., Mahmood A.K. National cyber security strategies for digital economy // International Conference on Research and Innovation in Information Systems, ICRIIS. 2017. P. 1–6.
2. Crabtree A. Enabling the new economic actor: Personal data regulation and the digital economy // Proceedings - 2016 IEEE International Conference on Cloud Engineering Workshops, IC2EW 2016. 2016. P. 124–129.
3. H. H. Tung, C. C. Cheng, Y. Y. Chen, Y. F. Chen S.H.H. and A.P.C. Binary Classification and Data Analysis for Modeling Calendar Anomalies in Financial Markets // 2016 7th International Conference on Cloud Computing and Big Data (CCBD). Macau, 2016. P. 116–121.
4. Trelewicz J.Q. Big Data and Big Money: The Role of Data in the Financial Sector // IT Prof. 2017. Vol. 19, № 3. P. 8–10.
5. Luvizan S.S., Nascimento P.T., Yu A. Big Data for innovation: The case of credit evaluation using mobile data analyzed by innovation ecosystem lens // PICMET 2016 - Portland International Conference on Management of Engineering and Technology: Technology Management For Social Innovation, Proceedings. 2017. P. 925–936.
6. M.U. Sapozhnikova, M.M. Gayanova, A.V. Nikonov A.M.V. Data mining algorithms of bank transaction as a part of antifraud system // Information technologies for intelligent decision making support. Ufa, 2017.
7. Lopez-Rojas E.A., Axelsson S. A review of computer simulation for fraud detection research in financial datasets // FTC 2016 - Proceedings of Future Technologies Conference. 2017. P. 932–935.
8. Big ideas are coming from using big data - Raconteur [Электронный ресурс]. 2014. URL: https://www.raconteur.net/technology/big-ideas-are-coming-from-using-big-data (дата обращения: 20.11.2017).
9. Steve Rosenbush. Visa Says Big Data Identifies Billions of Dollars in Fraud - CIO Journal [Электронный ресурс]. 2014. URL: https://blogs.wsj.com/cio/2013/03/11/visa-says-big-data-identifies-billions-of-dollars-in-fraud/ (дата обращения: 20.11.2017).
10. I.Piskunov. Anti-fraud systems and how it works [Электронный ресурс] 2017. URL: https://www.securitylab.ru/blog/personal/Informacionnaya_bezopasnost_v_detalyah/339929.php (дата обращения: 20.11.2017).
11. M.U. Sapozhnikova, M.M. Gayanova, A.V. Nikonov A.M.V. Data mining technologies in the problem of designing the bank transaction monitoring system // Computer Science and Information Technologies. Baden-Baden, 2017.
12. M.U. Sapozhnikova, M.M. Gayanova, A.V. Nikonov, A.M. Vulfin D.V.K. Anti-fraud system on the basis of data mining technologies // International Symposium on Signal Processing and Information Technology. Bilbao, 2017.
13. Abbad M., Abed J.M., Abbad M. The Development of E-Banking in Developing Countries in the Middle East. // J. Financ. Account. Manag. 2012. Vol. 3, № 2. P. 107–123.
14. Jarrett J.E. Internet Banking Development // J. Entrep. Organ. Manag. 2016. Vol. 5. P. 2–5.
15. Global Mass Payments, AP Software, B2B Payments Tipalti [Электронный ресурс]. URL: https://tipalti.com/ (дата обращения: 20.11.2017).
16. M. Fedotenko. Как защищают банки: разбираем устройство и принципы банковского антифрода - «Хакер» [Электронный ресурс]. 2017. URL: https://xakep.ru/2017/04/21/antifrod-1/ (дата обращения: 20.11.2017).
17. Cao Y., Li S., Wijmans E. (Cross-) Browser Fingerprinting via OS and Hardware Level Features // Proc. Netw. Distrib. Syst. Secur. Symp. 2017. № March.
18. Big Data [Электронный ресурс]. URL: https://en.wikipedia.org/wiki/Big_data (accessed: 20.11.2017).

19. Apache Hadoop [Электронный ресурс]. URL: http://hadoop.apache.org/ (accessed: 30.03.2018).
20. Apache Spark [Электронный ресурс]. URL: https://spark.apache.org/ (дата обращения: 30.03.2018).
21. ClickHouse [Электронный ресурс]. URL: https://clickhouse.yandex/ (дата обращения: 30.03.2018).
22. Elasticsearch [Электронный ресурс]. URL: https://www.elastic.co/products/elasticsearch (дата обращения: 30.03.2018).
23. Splunk [Электронный ресурс]. URL: https://www.splunk.com/ (дата обращения: 30.03.2018).
24. Apache Cassandra [Электронный ресурс]. URL: http://cassandra.apache.org/ (дата обращения: 30.03.2018).
25. Spark-sklearn [Электронный ресурс]. URL: https://github.com/databricks/spark-sklearn (дата обращения: 30.03.2018).
26. Sparkling Water [Электронный ресурс]. URL: https://www.h2o.ai/sparkling-water/ (дата обращения: 30.03.2018).
27. TensorFlowOnSpark [Электронный ресурс]. URL:https://github.com/yahoo/TensorFlowOnSpark (дата обращения: 30.03.2018).
28. Apache Zookeeper [Электронный ресурс]. URL: https://zookeeper.apache.org/ (дата обращения: 30.03.2018).
29. Apache Hive [Электронный ресурс]. URL: https://hive.apache.org/ (дата обращения: 30.03.2018).
30. Apache Sqoop [Электронный ресурс]. URL: http://sqoop.apache.org/ (дата обращения: 30.03.2018).
31. Apache Flume [Электронный ресурс]. URL: https://flume.apache.org/ (дата обращения: 30.03.2018).
32. Apache Oozie [Электронный ресурс]. URL: http://oozie.apache.org/ (дата обращения: 30.03.2018).
33. Cloudera Hue [Электронный ресурс]. URL: https://github.com/cloudera/hue (дата обращения: 30.03.2018).
34. Sentry [Электронный ресурс]. URL: https://sentry.io/welcome/ (дата обращения: 30.03.2018).
35. GitLab [Электронный ресурс]. URL: https://about.gitlab.com/ (дата обращения: 30.03.2018).