ПРИКАСПИЙСКИЙ ЖУРНАЛ

УПРАВЛЕНИЕ И ВЫСОКИЕ ТЕХНОЛОГИИ

2023
№4 (64)

ISSN 2074-1707

16+

ISSN 2074-1707

АСТРАХАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ В. Н. ТАТИЩЕВА

ПРИКАСПИЙСКИЙ ЖУРНАЛ:
управление и высокие технологии

НАУЧНО-ТЕХНИЧЕСКИЙ ЖУРНАЛ

2023
№ 4 (64)

Журнал включен в перечень рецензируемых научных изданий, рекомендованных ВАК России для публикации основных научных результатов диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук по следующим научным специальностям.

Группа специальностей 1.2 «Компьютерные науки и информатика»:

1.2.2 – Математическое моделирование, численные методы и комплексы программ (технические науки).

Группа специальностей 2.2 «Электроника, фотоника, приборостроение и связь»:

2.2.4 – Приборы и методы измерения (по видам измерений) (технические науки);

2.2.11 – Информационно-измерительные и управляющие системы (технические науки);

2.2.12 – Приборы, системы и изделия медицинского назначения (технические науки).

Группа специальностей 2.3 «Информационные технологии и телекоммуникации»:

2.3.1 – Системный анализ, управление и обработка информации (технические науки);

2.3.4 – Управление в организационных системах (технические науки);

2.3.5 – Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей (технические науки);

2.3.6 – Методы и системы защиты информации, информационная безопасность (технические науки).

Журнал входит в базу данных Ulrich’s Periodicals Directory.

Астрахань
Астраханский государственный университет имени В. Н. Татищева
2023

ББК 30
П75

Рекомендовано к печати редакционно-издательским советом
Астраханского государственного университета имени В. Н. Татищева

ПРИКАСПИЙСКИЙ ЖУРНАЛ: управление и высокие технологии

НАУЧНО-ТЕХНИЧЕСКИЙ ЖУРНАЛ

2023
№ 4 (64)

Журнал выходит 4 раза в год
Все материалы, поступающие в редколлегию журнала,
проходят независимое рецензирование

ISSN 2074-1707

ASTRAKHAN TATISHCHEV STATE UNIVERSITY

PRIKASPIYSKIY ZHURNAL:
Upravlenie i Vysokie Tekhnologii

CASPIAN JOURNAL:
Control and High Technologies

A SCIENTIFIC AND TECHNICAL JOURNAL

2023
No. 4 (64)

The journal is included in the list of the reviewed scientific journals recommended by VAK of Russia for the publication of the main scientific results of theses for the candidate of science degree, for the doctor of science degree on the following scientific specialties.

Group of specialties 1.2 "Computer science and informatics":

1.2.2 – Mathematical modelling, numerical methods and complexes of programms (technical sciences).

Group of specialties 2.2 "Electronics, photonics, instrument engineering and communication":

2.2.4 – Instruments and methods of measurement (by type of measurement) (technical sciences);

2.2.11 – Information-measuring and control systems (technical sciences);

2.2.12 – Medical devices, systems and products (technical sciences).

Group of specialties 2.3 "Information technologies and telecommunications":

2.3.1 – System analysis, information control and processing (technical sciences);

2.3.4 – Management in organizational systems (technical sciences);

2.3.5 – Mathematical software and software for computing systems, complexes and computer networks (technical sciences);

2.3.6 – Information security methods and systems, information security (technical sciences).

The journal is included into the database Ulrich's Periodicals Directory.

Astrakhan
Astrakhan Tatishchev State University
2023

Recommended by the Editorial and Publishing Board
of Astrakhan Tatishchev State University

**CASPIAN JOURNAL:
Control and High Technologies**

A SCIENTIFIC AND TECHNICAL JOURNAL

**2023
No. 4 (64)**

The journal is published four times a year
All materials that come to the Editorial Board of the journal
are subject to independent peer-review

СОДЕРЖАНИЕ

ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ

МЕТОДЫ И СИСТЕМЫ ЗАЩИТЫ ИНФОРМАЦИИ, ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ

УПРАВЛЕНИЕ В ОРГАНИЗАЦИОННЫХ СИСТЕМАХ

МАТЕМАТИЧЕСКОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ, КОМПЛЕКСОВ И КОМПЬЮТЕРНЫХ СЕТЕЙ

СИСТЕМНЫЙ АНАЛИЗ, УПРАВЛЕНИЕ И ОБРАБОТКА ИНФОРМАЦИИ

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ, ЧИСЛЕННЫЕ МЕТОДЫ И КОМПЛЕКСЫ ПРОГРАММ

CONTENTS

INFORMATICS, COMPUTER TECHNIQUE AND CONTROL

METHODS AND SYSTEMS OF INFORMATION PROTECTION, INFORMATION SECURITY

MANAGEMENT IN ORGANIZATIONAL SYSTEMS

MATHEMATICAL SOFTWARE AND SOFTWARE FOR COMPUTING MACHINES, COMPLEXES AND COMPUTER NETWORKS

SYSTEM ANALYSIS, CONTROL AND INFORMATION PROCESSING

MATHEMATICAL MODELLING, NUMERICAL METHODS AND PROGRAM SYSTEMS

ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ

МЕТОДЫ И СИСТЕМЫ ЗАЩИТЫ ИНФОРМАЦИИ, ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ

DOI 10.54398/20741707_2023_4_7
УДК 004.056.52

ПОРОГОВАЯ СХЕМА УПРАВЛЕНИЯ КЛЮЧАМИ ШИФРОВАНИЯ ДЛЯ ИНТЕЛЛЕКТУАЛЬНЫХ ТРАНСПОРТНЫХ СИСТЕМ НА БАЗЕ БЛОКЧЕЙН-ТЕХНОЛОГИИ

Носиров Зафаржон Амрулоевич, Финансовый университет при Правительстве Российской Федерации, 125167, Российская Федерация, г. Москва, проспект Ленинградский, 49/2,
аспирант, ORCID: 0000-0001-6858-1241, e-mail: nosirovzafar@outlook.com
Ажмухамедов Искандар Маратович, Астраханский государственный университет имени В. Н. Татищева, 414056, Российская Федерация, г. Астрахань, ул. Татищева, 20а,
доктор технических наук, профессор, ORCID: 0000-0001-9058-123X, e-mail: aim_agtu@mail.ru

В рамках статьи рассмотрена проблема обеспечения конфиденциальности данных, обрабатываемых в интеллектуальных транспортных системах, спроектированных на базе блокчейн-технологии. Проанализированы архитектурные подходы построения интеллектуальных транспортных систем и выявлены их недостатки. А также предложена пороговая схема управления секретами, которая помогает осуществлять корректное управление ключами шифрования между *n* заинтересованными объектами. В качестве основы для пороговой схемы управления секретами была выбрана схема Асмута – Блума. Представленная модель работы системы предоставляет гарантированный контроль над данными интеллектуальных транспортных систем, обеспечивает высокий уровень конфиденциальности и повышает отказоустойчивость системы.

Ключевые слова: схема управления ключами, децентрализация, конфиденциальность данных, схема разделения секрета

THRESHOLD ENCRYPTION KEY MANAGEMENT SCHEME FOR INTELLIGENT TRANSPORTATION SYSTEMS BASED ON BLOCKCHAIN TECHNOLOGIES

Nosirov Zafarzhon A., Financial University under the Government of the Russian Federation, 49/2 Leningradsky Ave., Moscow, 125167, Russian Federation,
graduate student, ORCID: 0000-0001-6858-1241, e-mail: nosirovzafar@outlook.com
Azhmukhamedov Iskandar M., Astrakhan Tatishchev State University, 20a Tatishchev St., Astrakhan, 414056, Russian Federation,
Doct. Sci. (Engineering), Professor, ORCID: 0000-0001-9058-123X, e-mail: aim_agtu@mail.ru

The article considers the problem of ensuring the confidentiality of data processed in intelligent transportation systems designed on the basis of blockchain technology. Architectural approaches of intelligent transportation systems construction are analyzed and their disadvantages are revealed. And a threshold scheme for secrets management is proposed, which helps to realize correct management of encryption keys between n interested entities. Asmuth-Bloom scheme has been chosen as the basis for the threshold secret management scheme. The presented model of system operation provides guaranteed control over intelligent transportation systems data, ensures a high level of confidentiality and improves the fault tolerance of the system.

Keywords: key management scheme, decentralization, data confidentiality, secret sharing scheme

Graphical annotation (Графическая аннотация)

ВВЕДЕНИЕ

В настоящее время Интернет вещей (IoT) [1, 2] переживает беспрецедентное развитие благодаря широкому распространению новых технологий, таких как большие данные и облачные вычисления [3]. В то же время развитие блокчейна [4, 5] также привнесло новые возможности и удобства. Например, применение блокчейна в интеллектуальных транспортных системах (ИТС) позволяет увеличить их уровень защищенности и надежности, а также достичь большей децентрализации. Следует отметить, что одним из важнейших вопросов промышленности и научных кругов являются вопросы информационной безопасности (ИБ), вызванные эволюцией ИТС в сторону централизации [6]. Отсутствие взаимного доверия между участниками является барьером для развития ИТС [7].

Использование блокчейна в ИТС не является панацеей, так как возникают вопросы ИБ, требующие детальной проработки. В частности, к таким вопросам относятся: аутентификация объектов ИТС, обеспечение конфиденциальности и целостности данных, а также надежности системы. В рамках исследования предлагается рассмотреть вопрос обеспечения конфиденциальности данных в ИТС, так как данный вопрос является малоисследованным. Это связано с тем, что на практике специалисты уделяют большое внимание модернизации конечных устройств, например, усовершенствованию RSU (Road side unit – это устройство, установленное на дорожной инфраструктуре) [8–10], а некоторые исследователи предлагают модернизировать существующие программные прошивки [11–13]. Лишь в малой части проанализированных работ предлагается существенно поменять устоявшиеся архитектурные подходы построения ИТС с учетом вопросов конфиденциальности данных [14–16].

Исходя из вышеизложенного, следует, что практически никто из исследователей не обращает внимание на вопросы обеспечения конфиденциальности данных, хранящихся в ИТС, поэтому можно считать данную тему исследования весьма актуальной.

Как известно, конфиденциальность данных достигается с помощью криптографии, а именно шифрования. Для шифрования информации можно использовать различные криптографические алгоритмы, позволяющие обеспечить безопасность данных как на стороне клиента, так и на стороне сервера. В общем случае шифрование можно разделить на симметричное и асимметричное. При этом важную роль как в симметричном, так и в асимметричном шифровании играет механизм управления ключами [17]. Управление ключами – это важнейший механизм шифрования, обеспечивающий безопасность ключа. Некорректное управление ключами может поставить под угрозу безопасность зашифрованных данных. Эффективными методами управления ключами являются протоколы обмена ключами [18], разделение секретов [19] и иерархическая генерация ключей [20]. В рамках данной работы в качестве методов управления ключами рассматриваются пороговые схемы разделения секретов, так как их применение является наиболее практичным и эффективным для ИТС.

Исходя из вышеизложенного, основной целью работы является разработка пороговой схемы управления ключами шифрования для ИТС на основе блокчейн-технологии.

АРХИТЕКТУРА ИНТЕЛЛЕКТУАЛЬНЫХ ТРАНСПОРТНЫХ СИСТЕМ

Для того чтобы понять, как устроены ИТС, необходимо рассмотреть их верхнеуровневую архитектуру. На рисунке 1 продемонстрирован упрощенный вариант классической архитектуры ИТС, в которой взаимодействуют транспортные средства через RSU, принадлежащие разным организациям.

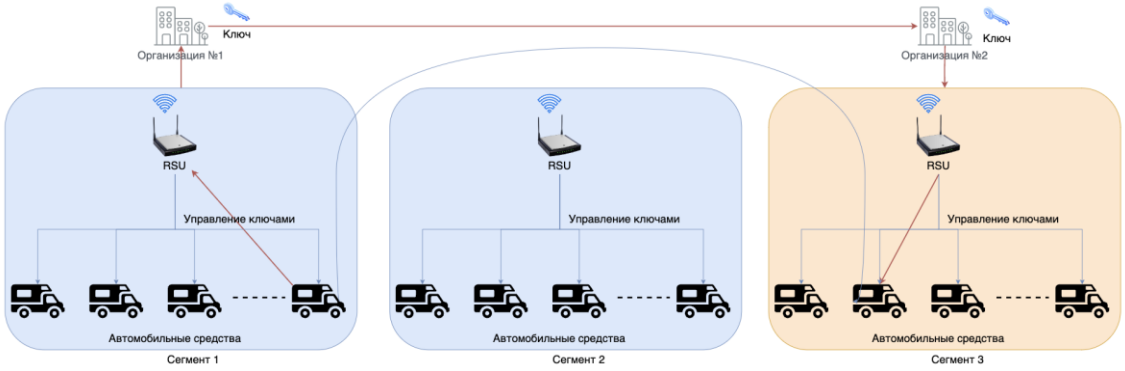


Рисунок 1 – Классическая архитектура ИТС

Ознакомившись с рисунком 1, становится ясно, что для взаимодействия двух транспортных средств, расположенных в разных сегментах, исходящему запросу (красные стрелки) необходимо преодолеть как минимум 4 узла. То есть исходящий запрос автомобильного средства из сегмента 1 сначала отправляется в RSU, затем RSU перенаправляет запрос в организацию № 1, далее осуществляется взаимодействие между двумя организациями. Получив запрос, организация № 2 перенаправляет запрос в RSU, который расположен наиболее близко к транспортному средству, которому был предназначен данный запрос. При использовании такого архитектурного подхода появляются дополнительные объекты, требующие повышенного внимания в части обеспечения информационной безопасности. К таким объектам относятся организации № 1 и № 2. Взлом систем одной из организаций может повлечь за собой сбой всей ИТС, а также может нарушить функционирование автомобильных средств. Поэтому можно утверждать, что классическая схема обмена данными ИТС подвержена различным угрозам ИБ и требуется переосмысление существующих архитектурных подходов построения ИТС.

В работе [21] предложена модифицированная архитектура ИТС, в которой применяется блокчейн-технология (рис. 2).

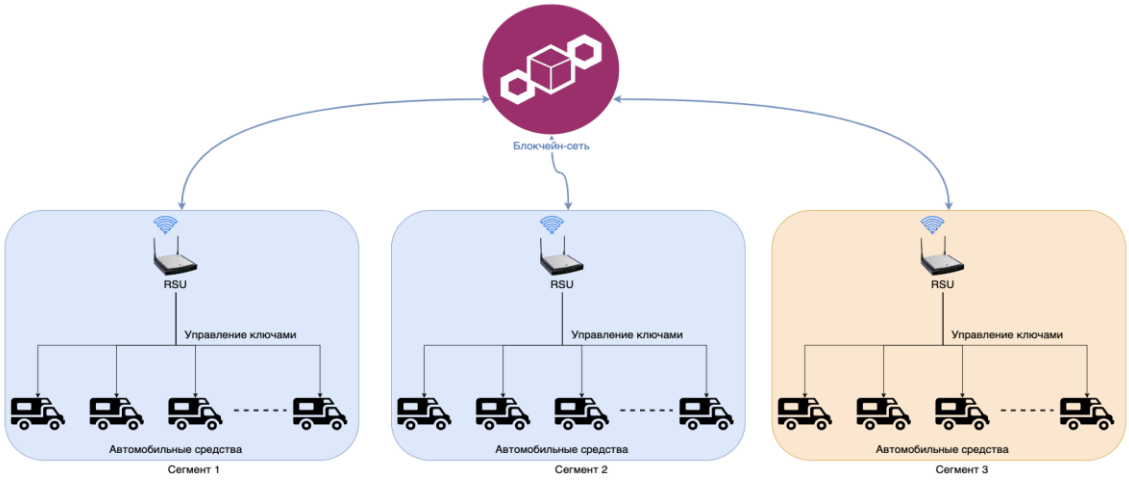


Рисунок 2 – Архитектура решения на основе блокчейн-сети

Результаты, полученные в вышеупомянутой работе, свидетельствуют о том, что применение блокчейн-технологии в ИТС позволяет упростить взаимодействие транспортных средств, находящихся в разных сегментах. Также важно отметить, что, используя преимущества блокчейн-технологии, можно гарантировать надежность связи и отказоустойчивость системы.

ПОРОГОВЫЕ СХЕМЫ РАЗДЕЛЕНИЯ СЕКРЕТА

Данный раздел посвящен анализу и выбору пороговой схемы разделения, которая ляжет в основу предлагаемого решения. Под пороговым разделением секретов в криптографии понимается схема, позволяющая разделить секрет s на n долей так, чтобы комбинация не менее чем из t долей могла восстановить секрет. В рамках данного исследования под секретом подразумевается ключ шифрования данных. Следует отметить, что, согласно результатам анализа, приведенным в работе [22], именно схемы Шамира и Асмута – Блума являются менее ресурсоемкими, поэтому было решено проанализировать их.

Схема разделения Шамира. Пороговая схема Шамира (t, n) построена вокруг концепции полиномиальной интерполяции. Если необходимо разделить секрет таким образом, чтобы восстановить его могли только *t* абонентов, то нужно «спрятать» его в формулу многочлена степени $(t-1)$. Восстанавливается этот многочлен по *t* точкам [23] согласно нижеприведенной формуле:

$$f(x) = a_0 + a_x x + a_2 x^2 + \cdots + a_{t-1} x^{t-1}. \quad (1)$$

Коэффициент *x* выбирается случайным образом, а секрет кодируется в виде константы a_0. Для того чтобы восстановить секрет (т. е. a_0), необходимо собрать не менее чем *t* хранителей. На основе полинома Лагранжа, представленного ниже, эти *t* хранителей могут восстановить полином $f(x)$, а значит, и восстановить секрет a_0:

$$L(x) = \sum_{j=0}^{t-1} y_j \cdot l_x(x). \quad (2)$$

На основе результатов работ Ади Шамира были предложены различные модификации схем разделения секретов, в которых применяются полиномы Лагранжа [24–26]. Также следует отметить, что схема Шамира используется в различных технологиях, таких как облачные вычисления [27, 28], и в алгоритмах обеспечения конфиденциальности [29].

Схема, основанная на китайской теореме об остатках. Схема разделения секрета [30], предложенная Асмутом и Блумом, основана на китайской теореме об остатке.

Если задан набор взаимно простых чисел $m_1, m_2, m_3, \ldots, m_n$, следующие линейные уравнения имеют единственное решение *M*, где $M = \prod_{i=1}^{n} m_i$:

$$\begin{cases} a_1 \bmod m_1 = x, \\ \ldots \\ a_n \bmod m_1 = x. \end{cases} \quad (3)$$

Единственное решение вычисляется по формуле:

$$x = [\sum_{i=1}^{n} a_i C^i (C_i^{-1} \bmod m_i)] \bmod M, \quad (4)$$

где $C_i = M/m_i$.

Вышерассмотренная схема является фундаментальной в криптографии, поэтому разделение секрета на основе китайской теоремы об остатках изучается по сей день. Последние исследования в области разделения секретов на основе данной схемы рассмотрены в работах [31–33]. Далее перейдем к основной части работы, а именно к разработке пороговой схемы управления ключами для ИТС.

ПОРОГОВАЯ СХЕМА УПРАВЛЕНИЯ КЛЮЧАМИ

Раздел посвящен разработке пороговой схемы управления ключами для ИТС на основе блокчейна. В частности, предложена модель работы системы и модифицированная архитектура обмена данными ИТС, а также проводится анализ оценки сложности вычислений пороговых схем.

На рисунке 3 представлена модель работы системы порогового управления ключами. Предполагается, что общие данные находятся в распоряжении *n* транспортных средств. Для облегчения использования и обмена данные хранятся в облаке. Однако хранение данных в открытом виде влечет за собой дополнительные угрозы информационной безопасности. Поэтому для обеспечения конфиденциальности автомобили генерируют симметричный ключ шифрования. Важно отметить, что в схеме безопасного обмена симметричный ключ шифрования делится на *n* частей и распределяется между автомобилями, участвующими на этапе разделения. Для того чтобы успешно восстановить ключ шифрования, необходимо собрать не менее *t* автомобилей.

Рисунок 3 – Модель работы системы

Таким образом, можно утверждать, что:

1) любой из *n* автомобилей имеет контроль над данными. В частности, любые *t* машин из *n* вместе могут восстановить симметричный ключ шифрования. Таким образом, они смогут расшифровать данные;

2) признание недействительными некоторых транспортных средств не приведет к невозможности восстановления ключа шифрования.

Управление ключами на основе схемы Шамира с использованием блокчейн-технологии может быть построено следующим образом:

1) генерация ключа шифрования: для обмена данными *Data* между *n* хранителями владельцу данных необходимо сгенерировать симметричный ключ шифрования AES. Ключ имеет вид $key \leftarrow \{0,1\}^l$. Здесь *l* – длина ключа шифрования, который может быть 256-битным;

2) выбор значения *t*: *n* хранителей совместно определяют минимальное количество легитимных хранителей, необходимых для восстановления секрета;

3) генерация полинома: владелец данных выбирает полином степени $t-1$ в соответствии с уравнением (1). Ключ кодируется в виде константы a_0, а остальные $t-1$ коэффициентов выбираются случайным образом;

4) формирование долей хранителей: сгенерировав полином, владелец данных для каждого *i* хранителя выбирает точку x_i и вычисляет y_i, далее распределяет их между хранителями, для того чтобы распределить ключ шифрования;

5) шифрование данных: после распределения симметричного ключа шифрования инициатор процесса разделения шифрует информацию *Data* с помощью ранее сгенерированного ключа и загружает зашифрованные данные *E* в облако. В облаке развернута наша блокчейн-сеть. Здесь $E = AES_{key}(Data)$;

6) расшифровка: с помощью полученной части хранитель вместе с другими $t-1$ участниками разделения секрета может полностью восстановить ключ. Затем хранители могут расшифровать зашифрованные данные *E*.

Рассмотрим пример схемы управления ключами. В данном примере участвуют 10 автомобильных средств, а $t = 4$. Выбранный полином представлен уравнением (5), а секрет имеет значение, равное 2023:

$$f(x) = 2023 + 8x + 25x^2 + 30x^3. \quad (5)$$

Данные 10 пар (x_i, y_i) распределяются между всеми автомобилями. В таблице приведены 10 пар (x_i, y_i), вычисленных на основе уравнения (5). Для облегчения понимания *x* задан в диапазоне от 2 до 11. Важно отметить, что на практике для сохранения безопасности значение x_i выбирается случайным образом.

Таблица – Результаты вычислений

x	2	3	4	5	6	7	8	9	10	11
y	2384	3082	4375	6438	9451	13594	19047	25990	34603	45066

Далее проверим, что любые 4 пары из таблицы могут быть использованы для восстановления секрета «2023». В примере для восстановления секрета выбраны следующие пары: (4, 4375), (5, 6438), (6, 9451) и (7, 13594).

$$L(x) = \sum_{j=0}^{t-1} y_j \cdot l_j(x) \Rightarrow L(0) = \sum_{j=0}^{t-1} y_j \cdot l_j(x) = y_1 \cdot \sum_{i=1, i\neq 1}^{4} \frac{-x_i}{x_1 - x_i} + y_2 \cdot \sum_{i=1, i\neq 2}^{4} \frac{-x_i}{x_2 - x_i} + y_3 \cdot \sum_{i=1, i\neq 3}^{4} \frac{-x_i}{x_3 - x_i} + y_4 \cdot \sum_{i=1, i\neq 4}^{4} \frac{-x_i}{x_4 - x_i} = 4375 \cdot \frac{5}{5-4} \cdot \frac{6}{6-4} \cdot \frac{7}{7-4} + 6438 \cdot \frac{4}{4-5} \cdot \frac{6}{6-5} \cdot \frac{7}{7-5} + 9451 \cdot \frac{4}{4-6} \cdot \frac{5}{5-6} \cdot \frac{7}{7-6} + 13594 \cdot \frac{5}{5-7} \cdot \frac{6}{6-7} \cdot \frac{4}{4-7} = 4375 \cdot 35 - 6438 \cdot 84 + 9451 \cdot 70 - 13594 \cdot 20 = 2023.$$

Для восстановления секрета на основе интерполяционного полинома нам было достаточно любых 4 пар (x_i, y_i). Кроме того, на рисунке 5 показаны три различных полинома, построенных на основе выбранного секрета «2023», полиномы y_1, y_2 и y_3 имеют вид $y_1 = 2023 + 10x + 45x^2 + 55x^3$, $y_2 = 2023 + 25x + 60x^2 + 81x^3$ и $y_3 = 2023 + 8x + 25x^2 + 30x^3$ соответственно.

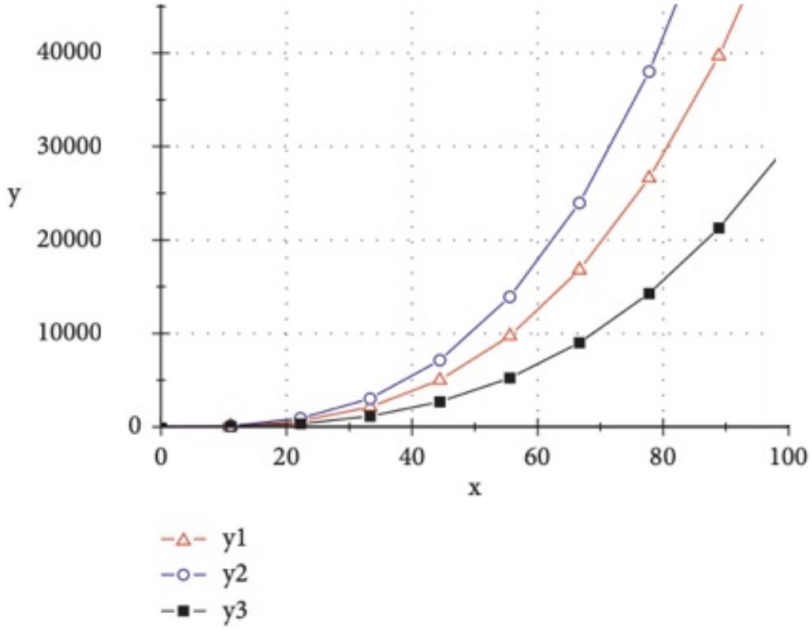


Рисунок 5 – Полиномы для секрета «2023»

Управление ключами на основе схемы Асмута – Блума с использованием блокчейн-технологии может быть реализовано следующим образом:

1) генерация ключей и выбор значения *t*: первые два шага идентичны шагам из схемы управления ключами на основе схемы Шамира. Владелец данных выбирает ключ AES определенной длины и значение для *t*;

2) выбор параметров: владелец данных выбирает *n* взаимо-простых чисел так, чтобы *m* таких были $(m_i, m_j) = 1$ для каждой пары m_i и m_j $(i \neq j)$. Затем владелец данных вычисляет произведение этих *n* простых чисел по формуле $M = \sum_{i=1}^{t} m_i$. Важно отметить, что выбранный секрет должен удовлетворять условиям $0 \leq key < m_1$;

3) формирование доли: для разделения секрета владелец данных выбирает случайное число *r* и вычисляет $S = \text{key} + r \cdot m_1$. Выбранное число *r* должно удовлетворять условиям $0 \leq r < M/m_1 - 1$;

4) распределение долей: для каждого *i*-го хранителя владелец данных распределяет $S_i = S \bmod m_i$;

5) шифрование и расшифровка: распределив симметричный ключ шифрования, инициатор процесса разделения шифрует информацию *Data* с помощью ранее сгенерированного ключа и загружает зашифрованную информацию в блокчейн-сеть. При этом для расшифровки необходимо участие не менее *t* заинтересованных хранителей. Для восстановления секрета им необходимо решить следующую систему уравнений:

$$\begin{cases} S_1 \bmod m_1 = S \\ \dots \\ S_t \bmod m_t = S \end{cases}. \quad (6)$$

Основываясь на китайской теореме об остатках, данное линейное уравнение имеет единственное решение:

$$S = \sum_{i=2}^{t} S_i \cdot C_i \cdot (C_i^{-1} \bmod m_i) \bmod M^*, \quad (7)$$

где $M^* = \prod_{i=2}^{t} m_i$ и $C_i = M^*/m_i$.

Для того чтобы проанализировать эффективность вышерассмотренных схем, необходимо провести оценку сложности этапа восстановления. На рисунке 6 показано сравнение двух ранее рассмотренных схем.

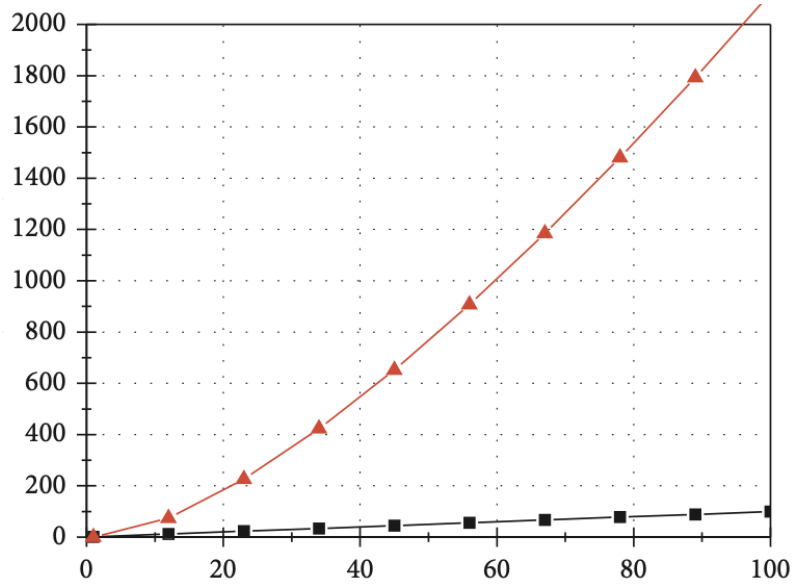


Рисунок 6 – Сравнение вычислительной сложности схемы Шамира (красная линия) и Асмута – Блума (черная линия)

Исходя из рисунка 6, становится ясно, что схема разделения секрета Асмута – Блума является наиболее эффективной в отличие от схемы Шамира. Поэтому в предлагаемом решении было решено использовать схему Асмута – Блума в качестве основы для порогового разделения ключа шифрования.

ЗАКЛЮЧЕНИЕ

В рамках данного исследования была рассмотрена проблема обеспечения конфиденциальности данных интеллектуальных транспортных систем, спроектированных на базе блокчейн-технологии. В частности, предложена пороговая схема управления ключами шифрования, которая помогает осуществлять корректное управление ключами между *n* заинтересованными объектами. Используя преимущества пороговых схем разделения секретов и блокчейн-технологии, мы смогли повысить уровень ИБ и отказоустойчивость процесса обмена данными в ИТС.

Список источников

1. Shen, J. Secure and efficient data sharing in dynamic vehicular networks / J. Shen, T. Zhou, J. Lai, P. Li and S. Moh // IEEE Internet of Things Journal. – 2020. – Vol. 7 (9). – P. 8208–8217.

2. Su, Y. Deep learning methods in internet of medical things for valvular heart disease screening system / Y. Su, T. Ding, M. Chen // IEEE Internet of Things Journal. – 2021. – P. 1–13.

3. Liang, L. Efficient and secure decision Tree classification for cloud-assisted online diagnosis services / J. Liang, Z. Qin, S. Xiao, L. Ou, L. Lin // IEEE Transactions on Dependable and Secure Computing, 2021. – Vol. 18 (4). – P. 1632–1644.

4. Tan, H. Practical homomorphic authentication in cloud-assisted vanets with blockchain-based healthcare monitoring for pandemic control / H. Tan, P. Kim, and I. Chung // Electronics. – 2020. – Vol. 9 (10). – P. 1683–1670.

5. Носиров З. А. Анализ блокчейн-технологии: основы архитектуры, примеры использования, перспективы развития, проблемы и недостатки / З. А. Носиров, В. М. Фомичев // Системы управления, связи и безопасности. – 2021. – № 2. – URL: https://cyberleninka.ru/article/n/analiz-blokcheyn-tehnologii-osnovy-arhitektury-primery-ispolzovaniya-perspektivy-razvitiya-problemy-i-nedostatki (дата обращения: 20.09.2023).

6. Zhang, J. Data driven intelligent transportation systems: a survey / J. Zhang, F. Wang, K. Wang, W. Lin, X. Xu, C. Chen // IEEE Transactions on Intelligent Transportation Systems. – 2020. – Vol. 12. – P. 1624–1639.

7. Singh, P. Blockchain and AI technology convergence: Applications in transportation systems. / P. Singh // Vehicular Communications. – 2022. – P. 521–529.

8. Gamboa-Rosales, A. Visualizing the intellectual structure and evolution of intelligent transportation systems: A systematic analysis of research themes and trends / A. Gamboa-Rosales, N. Karina // Sustainability 12.21. – 2020. – P. 8759–8763.

9. Waqas, M. Authentication of Vehicles and Road Side Units in Intelligent Transportation System / M. Waqas, M. Muhammad // Computers, Materials & Continua. – 2020. – P. 64.

10. Kumar, J. A privacy-preserving-based secure framework using blockchain-enabled deep-learning in cooperative intelligent transport system / J. Kumar, A. Randhir // IEEE Transactions on Intelligent Transportation Systems. – 2021. – Vol. 23 (9). – P. 16492–16503.

11. Lăzăroiu, A. Connected and autonomous vehicle mobility: Socially disruptive technologies, networked transport systems, and big data algorithmic analytics / A. Lăzăroiu, George, V. Machová, and J. Kucera // Contemporary Readings in Law and Social Justice. – 2020. – Vol. 12 (2). – P. 61–69.

12. Akhmatova, A. Integrating quality management systems (TQM) in the digital age of intelligent transportation systems industry 4.0 / A. Akhmatova, Malika-Sofi // Transportation Research Procedia. – 2022. – Vol. 6. – P. 1512–1520.

13. Ashraf, A. Novel deep learning-enabled LSTM autoencoder architecture for discovering anomalous events from intelligent transportation systems / A. Ashraf, Javed // IEEE Transactions on Intelligent Transportation Systems. – 2020. – Vol. 22 (7). – P. 4507–4518.

14. Priyanka, E. Application of integrated IoT framework to water pipeline transportation system in smart cities / E. Priyanka // Intelligence in Big Data Technologies – Beyond the Hype : proceedings of ICBDCC. – 2021. – P. 45–52.

15. Kripak, M. Analytical support for effective functioning of intelligent manufacturing and transport systems / M. Kripak, S. Palkina, A. Seliverstov // IOP Conference Series: Materials Science and Engineering, 2020. – Vol. 709 (3). – P. 32–40.

16. Lieberman, S. Comparison of intelligent transportation systems based on biocybernetic vehicle control systems / S. Lieberman, P. Klachek, S. Korjagin. // Transportation research procedia, 2020. – Vol. 50. – P. 355–362.

17. Zhou, T. Key agreement protocol with dynamic property for vanets / T. Zhou, H. Yang, J. Shen // Journal of Cryptologic Research. – 2020. – Vol. 7 (3). – P. 1–14.

18. Faz-Hernandez, A. A faster software implementation of the supersingular isogeny Diffie-Hellman key exchange protocol / A. Faz-Hernandez, J. Lopez, E. Ochoa-Jimenez, F. Rodrıguez-Henrıquez // IEEE Transactions on Computers. – 2017. – Vol. 67. – P. 1622–1636.

19. Shen, J. A novel Latin- square-based secret sharing for m2m communications / J. Shen, T. Zhou, X. Liu, Y. Chang // IEEE Transactions on Industrial Informatics. – 2018. – Vol. 14 (8). – P. 3659–3668.

20. Albakri, A. Hierarchical key management scheme with probabilistic security in a wireless sensor network (WSN) / A. Albakri, L. Harn, S. Song // Security and Communication Networks, 2019. – Vol. 19. – P. 11–17.

21. Zhou, A. Threshold key management scheme for blockchain-based intelligent transportation systems / A. Zhou, A. Tianqi // Security and Communication Networks. – 2021. – P. 1–8.

22. Носиров, З. А. Анализ криптографических схем разделения секрета для резервного хранения ключевой информации / З. А. Носиров, О. В. Щербинина // Прикаспийский журнал: управление и высокие технологии. – 2019. – № 2 (46). – С. 126–134. – EDN FFNPKK.

23. Shamir, A. How to share a secret / A. Shamir // Communications of the ACM, 1979. – Vol. 22 (11). – P. 612–613.

24. Dawson, E. The breadth of Shamir's secret sharing scheme / E. Dawson, D. Donovan // Computers & Security. – 1974. – Vol. 13 (1). – P. 69–78.

25. Benzekki, K. A verifiable secret sharing approach for secure multi-cloud storage / K. Benzekki, A. Fergougui, A. Elalaoui // Proceedings of the International Symposium on Ubiquitous Networking. – 2019. – P. 225–234.

26. Arbogast, J. Parallelizing shamir's secret sharing algorithm / J. Arbogast, I. Sumner, O. Lam // Journal of Computing Sciences in Colleges. – 2018. – Vol. 33. – P. 12–18.

27. Pundkar, N. Cloud computing security in multi-clouds using shamir's secret sharing scheme / N. Pundkar, N. Shekokar // Proceedings of the 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT). – 2016. – P. 392–395.

28. Zhou, T. Movie recommendation system employing the user-based in cloud computing / T. Zhou, L. Chen, J. Shen // Proceedings of the IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC). – 2017. – P. 46–50.

29. Li, Q. A privacy-preserving asynchronous averaging algorithm based on shamir's secret sharing / Q. Li, M. Christensen // Proceedings of the 2019 27th European Signal Processing Conference (EUSIPCO). – 2019. – P. 1–5.

30. Asmuth, C. A modular approach to key safeguarding / C. Asmuth, J. Bloom // IEEE Transactions on Information Theory. – 1983. – Vol. 29 (2). – P. 208–210.

31. Ersoy, O. Homomorphic extensions of CRT-based secret sharing/ O. Ersoy, B. Pedersen, E. Anarim // Discrete Applied Mathematics. – 2020. – Vol. 285. – P. 317–329.

32. Jia, X. A new threshold changeable secret sharing scheme based on the Chinese remainder theorem / X. Jia, D. Wang, D. Nie, X. Luo, J. Sun // Information Sciences. – 2019. – Vol. 473. – P. 13–30.

33. Ersoy, O. A CRT-based verifiable secret sharing scheme secure against unbounded adversaries / O. Ersoy, T. Pedersen, K. Kaya, A. Selcuk, E. Anarim //Security and Communication Networks. – 2016. – Vol. 9 (17). – P. 4416–4427.

References

1. Shen, J. et al. Secure and efficient data sharing in dynamic vehicular networks. *IEEE Internet of Things Journal*, 2020, vol. 7 (9), pp. 8208–8217.

2. Su, Y. et al. Deep learning methods in internet of medical things for valvular heart disease screening system. *IEEE Internet of Things Journal*, 2021, pp. 1–13.

3. Liang, L. et al. Efficient and secure decision Tree classification for cloud-assisted online diagnosis services. *IEEE Transactions on Dependable and Secure Computing*, 2021, vol. 18 (4), pp. 1632–1644.

4. Tan, H. et al. Practical homomorphic authentication in cloud-assisted vanets with blockchain-based healthcare monitoring for pandemic control. *Electronics*, 2020, vol. 9 (10), pp. 1683–1670.

5. Nosirov, Z. A., Fomichev, V. M. Analysis of Blockchain Technology: Architectural Basics, Application Examples, Future Trends, Problems and Disadvantages. *Systems of Control, Communication and Security*, 2021, no. 2. URL: https://cyberleninka.ru/article/n/analiz-blokcheyn-tehnologii-osnovy-arhitektury-primery-ispolzovaniya-perspektivy-razvitiya-problemy-i-nedostatki (дата обращения: 20.09.2023).

6. Zhang, J. et al. Data driven intelligent transportation systems: a survey. *IEEE Transactions on Intelligent Transportation Systems*, 2020, vol. 12, pp. 1624–1639.

7. Singh, P. Blockchain and AI technology convergence: Applications in transportation systems. *Vehicular Communications*, 2022, pp. 521–529.

8. Gamboa-Rosales, A. et al. Visualizing the intellectual structure and evolution of intelligent transportation systems: A systematic analysis of research themes and trends. *Sustainability*, 2020, pp. 8759–8763.

9. Waqas, M. et al. Authentication of Vehicles and Road Side Units in Intelligent Transportation System. *Computers, Materials & Continua*, 2020, p. 64.

10. Kumar, J. et al. A privacy-preserving-based secure framework using blockchain-enabled deep-learning in cooperative intelligent transport system. *IEEE Transactions on Intelligent Transportation Systems*, 2021, vol. 23 (9), pp. 16492–16503.

11. Lazaroiu, A. et al. Connected and autonomous vehicle mobility: Socially disruptive technologies, networked transport systems, and big data algorithmic analytics. *Contemporary Readings in Law and Social Justice*, 2020, vol. 12 (2), pp. 61–69.

12. Akhmatova, A. Integrating quality management systems (TQM) in the digital age of intelligent transportation systems industry 4.0 / A. Akhmatova, Malika-Sofi. *Transportation Research Procedia*, 2022, vol. 6, pp. 1512–1520.

13. Ashraf, A. et al. Novel deep learning-enabled LSTM autoencoder architecture for discovering anomalous events from intelligent transportation systems. *IEEE Transactions on Intelligent Transportation Systems*, 2020, vol. 22 (7), pp. 4507–4518.

14. Priyanka, E. Application of integrated IoT framework to water pipeline transportation system in smart cities. *Intelligence in Big Data Technologies—Beyond the Hype: Proceedings of ICBDCC*, 2021, pp. 45–52.

15. Kripak, M. et al. Analytical support for effective functioning of intelligent manufacturing and transport systems. *IOP Conference Series: Materials Science and Engineering*, 2020, vol. 709 (3), pp. 32–40.

16. Lieberman, S. et al. Comparison of intelligent transportation systems based on biocybernetic vehicle control systems. *Transportation research procedia*, 2020, vol. 50, pp. 355–362.

17. Zhou, T. et al. Key agreement protocol with dynamic property for vanets. *Journal of Cryptologic Research*, 2020, vol. 7 (3), pp. 1–14.

18. Faz-Hernandez, A. et al. A faster software implementation of the supersingular isogeny Diffie-Hellman key exchange protocol. *Transactions on Computers*, 2017, vol. 67, pp. 1622–1636.

19. Shen, J. et al. A novel Latin- square-based secret sharing for m2m communications. *IEEE Transactions on Industrial Informatics*, 2018, vol. 14 (8), pp. 3659–3668.

20. Albakri, A. et al. Hierarchical key management scheme with probabilistic security in a wireless sensor network (WSN). *Security and Communication Networks*, 2019, vol. 19, pp. 11–17.

21. Zhou, A. et al. Threshold key management scheme for blockchain-based intelligent transportation systems. *Security and Communication Networks*, 2021, pp. 1–8.

22. Nosirov, Z. A. Analysis of cryptographic secret sharing schemes for backing up key information. *Caspian Journal: Control and High Technologies*, 2019, vol. 2 (46), pp. 126–134.

23. Shamir, A. How to share a secret. *Communications of the ACM*, 1979, vol. 22 (11), pp. 612–613.

24. Dawson, E. et al. The breadth of Shamir's secret sharing. *Computers & Security*, 1974, vol. 13 (1), pp. 69–78.

25. Benzekki, K. et al. A verifiable secret sharing approach for secure multi-cloud storage. *Proceedings of the International Symposium on Ubiquitous Networking*, 2019, pp. 225–234.

26. Arbogast, J. et al. Parallelizing shamir's secret sharing algorithm. *Journal of Computing Sciences in Colleges*, 2018, vol. 33, pp. 12–18.

27. Pundkar, N. et al. Cloud computing security in multi-clouds using shamir's secret sharing scheme. *Proceedings of the 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 2016, pp. 392–395.

28. Zhou, T. et al. Movie recommendation system employing the user-based in cloud computing. *Proceedings of the IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, 2017, pp. 46–50.

29. Li, Q. et al. A privacy-preserving asynchronous averaging algorithm based on shamir's secret sharing. *Proceedings of the 2019 27th European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

30. Asmuth, C. et al. A modular approach to key safeguarding. *IEEE Transactions on Information Theory*, 1983, vol. 29 (2), pp. 208–210.

31. Ersoy, O. et al. Homomorphic extensions of CRT-based secret sharing. *Applied Mathematics*, 2020, vol. 285, pp. 317–329.

32. Jia, X. et al. A new threshold changeable secret sharing scheme based on the Chinese remainder theorem. *Information Sciences*, 2019, vol. 473, pp. 13–30.

33. Ersoy, O. et al. A CRT-based verifiable secret sharing scheme secure against unbounded adversaries. *Security and Communication Networks*, 2016, vol. 9 (17), pp. 4416–4427.

Статья поступила в редакцию 25.09.2023; одобрена после рецензирования 29.09.2023; принята к публикации 12.10.2023.

The article was submitted 25.09.2023; approved after reviewing 29.09.2023; accepted for publication 12.10.2023.

DOI 10.54398/20741707_2023_4_16
УДК 004.896

АВТОМАТИЗАЦИЯ ОЦЕНКИ УЯЗВИМОСТЕЙ ПРОГРАММНЫХ, ПРОГРАММНО-АППАРАТНЫХ СРЕДСТВ В ЦЕЛЕВОЙ ИНФОРМАЦИОННОЙ СИСТЕМЕ

Жуков Вадим Геннадьевич, Сибирский государственный университет науки и технологий, 660037, Российская Федерация, г. Красноярск, пр. им. газ. «Красноярский рабочий», 31,
кандидат технических наук, доцент, ORCID: 0000-0002-7933-6820, e-mail: zhukov@sibsau.ru
Селигеев Сергей Викторович, Сибирский государственный университет науки и технологий, 660037, Российская Федерация, г. Красноярск, пр. им. газ. «Красноярский рабочий», 31,
студент, ORCID: 0009-0002-7762-0382, e-mail: seligeevsergei@gmail.com

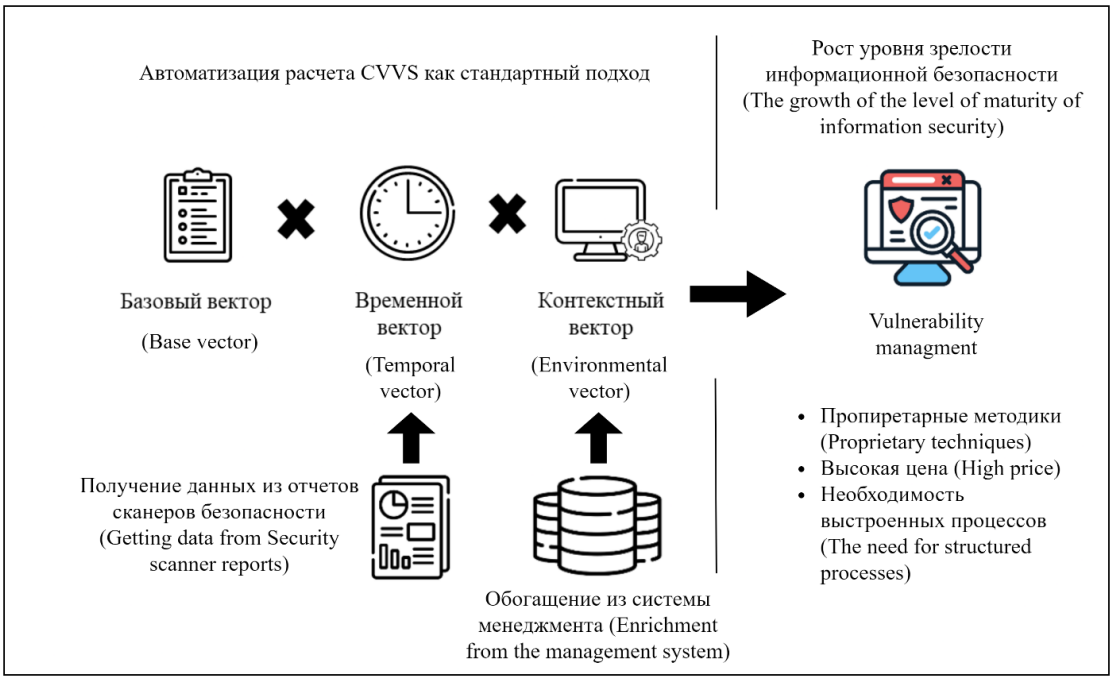
В работе рассматривается автоматизация расчета значения оценки критичности уязвимостей программных и программно-аппаратных средств в целевой информационной системе по методике CVSS (совокупное значение, полученное по базовому, временному и контекстному векторам) средствами свободно-распространяемого программного обеспечения – сканер безопасности OpenVAS и система инвентаризации активов GLPI. Предлагается в системе инвентаризации активов карту каждого узла ИТ-инфраструктуры (или только системно-значимых, критичных) расширить специальными дополнительными информационными полями, которые будут инициализироваться при вводе в эксплуатацию и поддерживаться в актуальности в процессе их эксплуатации. Значения этих полей будут использоваться для автоматической инициализации контекстных метрик. Для автоматической инициализации временных метрик предлагается анализ и извлечение необходимых значений (там, где они доступны) из отчетов сканеров безопасности. Значения оценок, полученных по базовому, временному и контекстному векторам, позволят более объективно оценивать критичность уязвимостей в целевой ИТ-инфраструктуры и могут быть использованы как базис в других методиках оценки. Опыт применения предложенного решения позволит специалистам в будущем лучше понимать принципы работы решений класса Vulnerability Management в рамках используемой ими системы оценки критичности уязвимостей и тем самым более эффективно и оперативно выявлять, анализировать и оперативно устранять уязвимости программных и программно-аппаратных средств в целевой информационной системе.
Ключевые слова: информационная безопасность, автоматизация, уязвимость, CVSS

AUTOMATION OF VULNERABILITY ASSESSMENT OF SOFTWARE, HARDWARE AND SOFTWARE IN THE TARGET INFORMATION SYSTEM

Zhukov G. Vadim, Siberian State University of Science and Technology, 31 Krasnoyarsky Rabochy Ave., Krasnoyarsk, 660037, Russian Federation,
Cand. Sci. (Engineering), Associate Professor, ORCID: 0000-0002-7933-6820, e-mail: zhukov@sibsau.ru
Seligeev V. Sergey, Siberian State University of Science and Technology, 31 Krasnoyarsky Rabochy Ave., Krasnoyarsk, 660037, Russian Federation,
student, ORCID: 0009-0002-7762-0382, e-mail: seligeevsergei@gmail.com

The paper considers the automation of calculating the value of the assessment of the criticality of software and firmware vulnerabilities in the target information system using the CVSS method (cumulative value obtained from the base, temporal and context vectors) using freely distributed software – the OpenVAS security scanner and the GLPI asset inventory system. It is proposed in the asset inventory system to expand the map of each IT infrastructure node (or only systemically significant, critical ones) with special additional information fields that will be initialized during commissioning and kept up to date during their operation. The values of these fields will be used to automatically initialize the context metrics. For automatic initialization of temporal metrics, it is proposed to analyze and extract the necessary values (where they are available) from reports of security scanners. The assessment values obtained by the base, time and context vectors will allow a more objective assessment of the criticality of vulnerabilities in the target IT infrastructure and can be used as a basis in other assessment methods. The experience of applying the proposed solution will allow specialists in the future to better understand the principles of operation of Vulnerability Management class solutions within the framework of the vulnerability criticality assessment system they use and, thereby, more effectively and quickly identify, analyze and promptly eliminate software and firmware vulnerabilities in the target information system.
Keywords: information security, automation, vulnerability, CVSS

Graphical annotation (Графическая аннотация)



ВВЕДЕНИЕ

Автоматизация бизнес-процессов является неотъемлемой составляющей операционной и стратегической деятельности компаний. Под автоматизацией бизнес-процессов понимается их реализация и контроль за ними средствами информационной системы. Примерами такой автоматизации являются системы автоматизации взаимодействия с потребителями, системы автоматического управления складами, поставками, производством, бухгалтерией, человеческими ресурсами и т. п. Таким образом, сама возможность компании функционировать зависит от доступности, стабильности, безопасности программных и программно-аппаратных средств автоматизации и данных, необходимых для их корректной работы. Высокая скорость разработки, сложность разрабатываемых решений, большое количество разработчиков, ограниченное время на тестирование и другие причины приводят к ошибкам и уязвимостям в коде программных и программно-аппаратных средств, а некорректные их настройки, настройки по умолчанию, приводят к уязвимостям конфигурации. Наличие уязвимостей создает угрозу их эксплуатации и тем самым возможность нарушить или прекратить работу программных и программно-аппаратных средств. Данная проблема усугубляется большим количеством уязвимостей [1] – общее количество уязвимостей даже в небольших инфраструктурах может исчисляться сотнями и тысячами, а критичных – десятками. Для обнаружения, анализа и оперативного устранения уязвимостей профильные специалисты применяют специализированные решения класса Vulnerability Management (VM). Особенностью VM-решений является то, что их производители разрабатывают, реализуют и поддерживают собственные проприетарные методики оценки и приоритизации уязвимостей, например, трендовые уязвимости от Positive Technologies [2], рейтинг VPR от Tenable [3], Qualys TruRisk [4]. Эти методики действительно позволяют эффективно оценивать и приоритизировать обнаруженные уязвимости за счет большого количества, затраченного человеко-часов специалистами производителя, чей основной задачей является выявление, исследование и структурирование данных об уязвимостях. Фактически приобретение VM-решений представляет из себя передачу решения задачи определения критичности уязвимостей и/или приоритизации на аутсорсинг поставщику этого решения. Но далеко не все могут позволить приобретение данных решений, а автоматизировать процесс обнаружения, анализа и оперативного устранения уязвимостей необходимо в том или ином виде любой компании, в которой существует хоть какая-то автоматизация. Например, опрос Positive Technologies «Как изменилась работа с уязвимостями в 2022 году», участие в котором принимали 200000 человек, говорит о том, что 52 % опрошенных доверяют и используют оценку CVSS [5–7]. Де-факто применение CVSS является общим знаменателем всех VM-решений. CVSS предоставляет возможность оценить уязвимость, но для полнофункционального применения и получения объективных значений оценки для каждой уязвимости необходима информация для инициализации всех метрик – базовых, временных и контекстных. VM-решения поставляют оценки только по базовому вектору в обязательном порядке

и опционально временные для некоторых уязвимостей, инициализацию контекстных метрик предлагается возложить на специалистов на местах. Без приземления на целевую инфраструктуру (инициализация контекстных метрик), без автоматизации полного расчета CVSS по всем метрикам использование CVSS будет малоэффективным, и, наоборот, автоматизация расчета полной оценки CVSS (хотя бы для некоторых уязвимостей) является актуальной задачей, решение которой будет достаточным для принятия решения о приоритизации уязвимостей в большинстве компаний и с повышением уровня зрелости информационной безопасности необходимым для понимания принципов работы и эффективного применения VM-решений.

НЕОБХОДИМОСТЬ УЧЕТА И КЛАССИФИКАЦИИ АКТИВОВ

Так как даже в небольших инфраструктурах количество активов и, соответственно, программного обеспечения на них может исчисляться десятками и сотнями, то инициализация контекстных метрик уязвимостей для всего перечня программных и программно-аппаратных средств может занять значительное количество ресурсов (человеко-часов), который де-факто ограничен, что может негативно повлиять на процесс управления уязвимостями в целом. Также активы могут добавляться или изменять положение в инфраструктуре, что создает необходимость переинициализации метрик контекстного вектора [8] для каждого из них, что снова возвращает к проблеме ограниченного ресурса.

Поэтому в качестве компромисса автоматический расчет полной оценки CVSS явно реализуется для уязвимостей на критичных активах, а однотипные активы сводятся в группы, контекстные метрики которых будут одинаковы. Данные ограничения позволят значительно сократить время, затрачиваемое на инициализацию метрик контекстного вектора уязвимостей для различных активов.

Наиболее ценным фактором при определении критичности актива являются убытки, как финансовые, так и репутационные, которые организация понесет в случае успешной эксплуатации уязвимостей и в результате нанесенного ущерба. В рамках предложенного подхода критичность актива будет являться основой для коррекции оценки, что возможно за счет особенностей логики расчета полной оценки CVSS.

ОСОБЕННОСТИ ЛОГИКИ РАСЧЕТА CVSS

Критичность актива должна определяться владельцем актива (не специалистом по информационной безопасности) с позиции влияния актива на бизнес-процесс или функцию в организации. Примем, что критичность актива в ИТ-инфраструктуре будет определять и отражаться в значениях метрик требований безопасности (Security Requirements) контекстного вектора: требования конфиденциальности (Confidentiality Requirement, CR), целостности (Integrity Requirement, IR), доступности (Availability Requirement, AR). Данные значения определяются специалистом по информационной безопасности на основании установленной критичности актива. В будущем с повышением уровня зрелости добавляется возможность учитывать дополнительные параметры, получать большее количество данных. Фактически, имея лишь значения метрик требований безопасности (CR, IR, AR) и приравняв значения модифицированных базовых метрик (Modified Base Metric) в контекстном векторе к значениям метрик базового вектора, становится возможным скорректировать оценку в соответствии с установленной критичностью актива. Если инициализировать значения метрик требований безопасности значением по умолчанию Medium, а метрики временного вектора – Not Defined, то оценка по контекстному вектору будет численно равна оценке по базовому вектору (рис. 1 и 2) – приведённый вектор в данном случае не относится к конкретной уязвимости и служит лишь для демонстрации логики рассуждений.

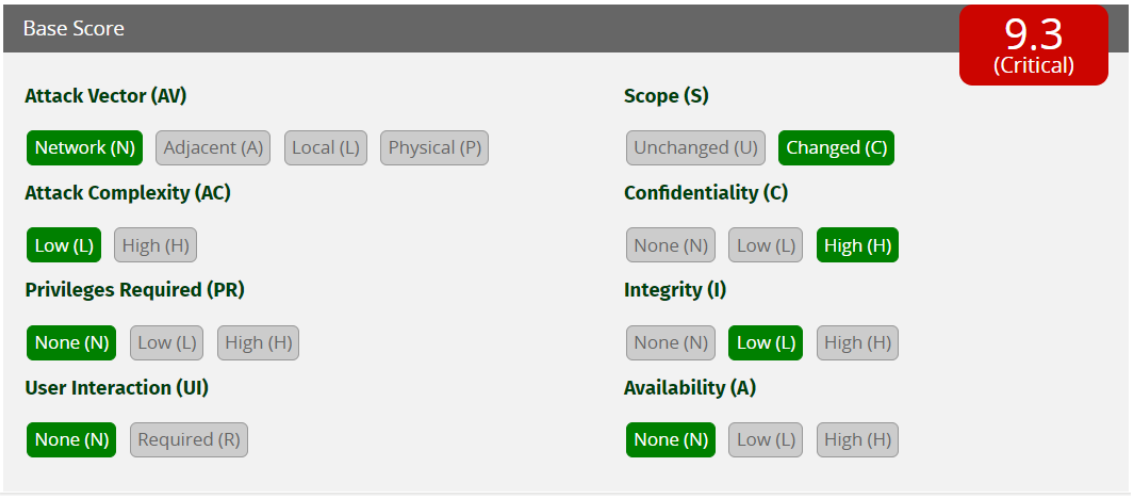


Рисунок 1 – Оценка по базовому вектору

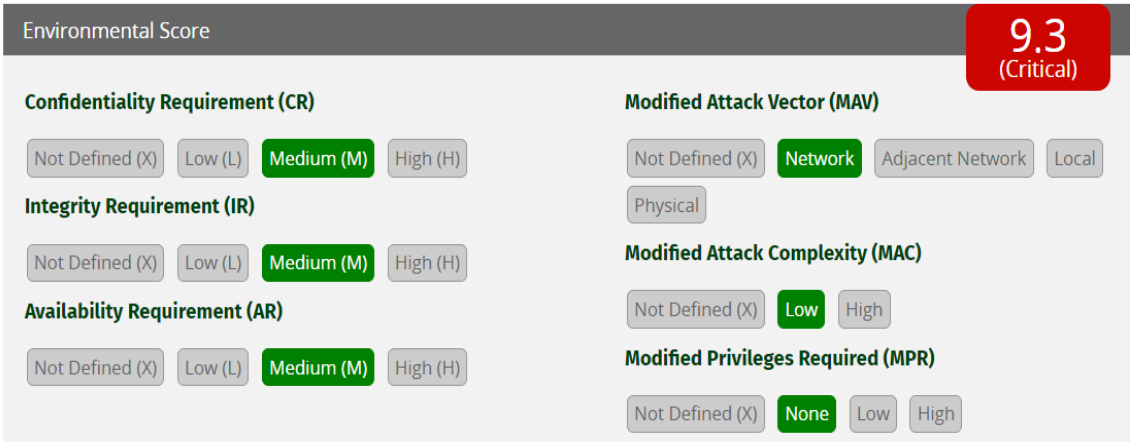


Рисунок 2 – Оценка по контекстному вектору при значениях по умолчанию

Манипуляция значениями метрик CR, IR и AR в калькуляторе CVSS v3.1 может довольно значительно повлиять на оценку уязвимости (от 6,9 (Medium) до 10 (Critical)), что изображено на рисунках 3 и 4.

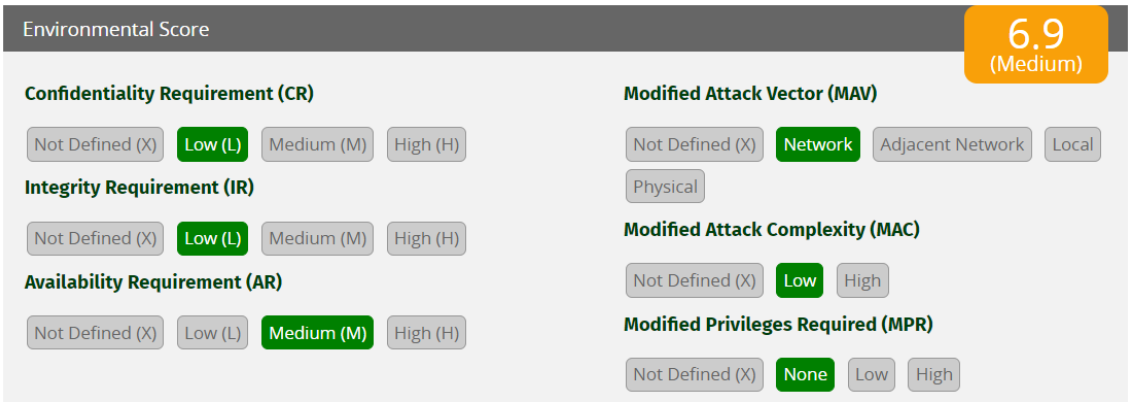


Рисунок 3 – Условная низкая критичность актива

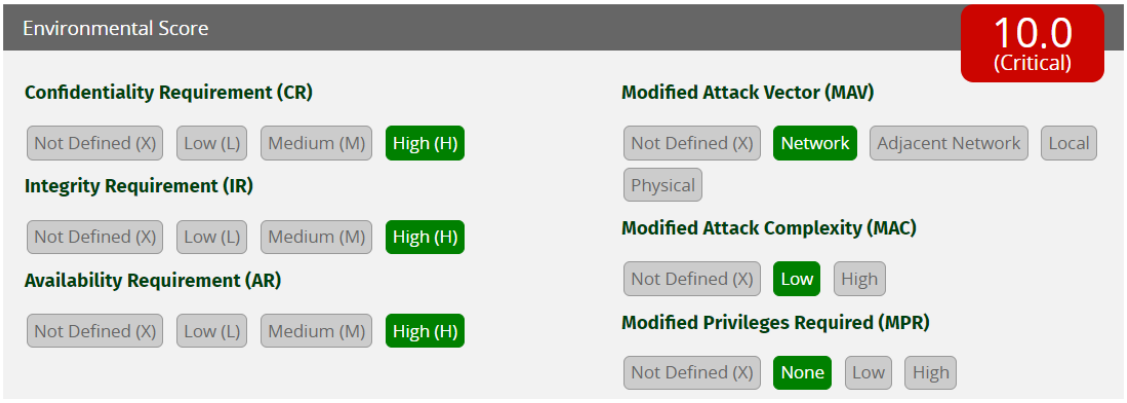


Рисунок 4 – Условная высокая критичность актива

Приведенные примеры, включают в себя оценки по всем трем векторам CVSS, что фактически является полной оценкой CVSS [5–7]. Расширить количество известных параметров можно за счет модернизации системы инвентаризации активов и использования отчетов о сканировании.

ПРИМЕНЯЕМОЕ ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ

Основной идеей предлагаемого подхода является построение алгоритма расчета полной оценки CVSS [5–7] с использованием информации из отчетов сканеров безопасности и системы

инвентаризации активов. Для получения значений временного вектора анализируются отчеты сканеров безопасности, а для получения значений метрик контекстного вектора используются дополнительные информационные поля, которые содержат конкретные значения для метрик контекстного вектора для групп или конкретных активов. В итоге общий алгоритм получения оценки, с учетом особенностей логики расчета CVSS, разобранных ранее, можно увидеть на рисунке 5.

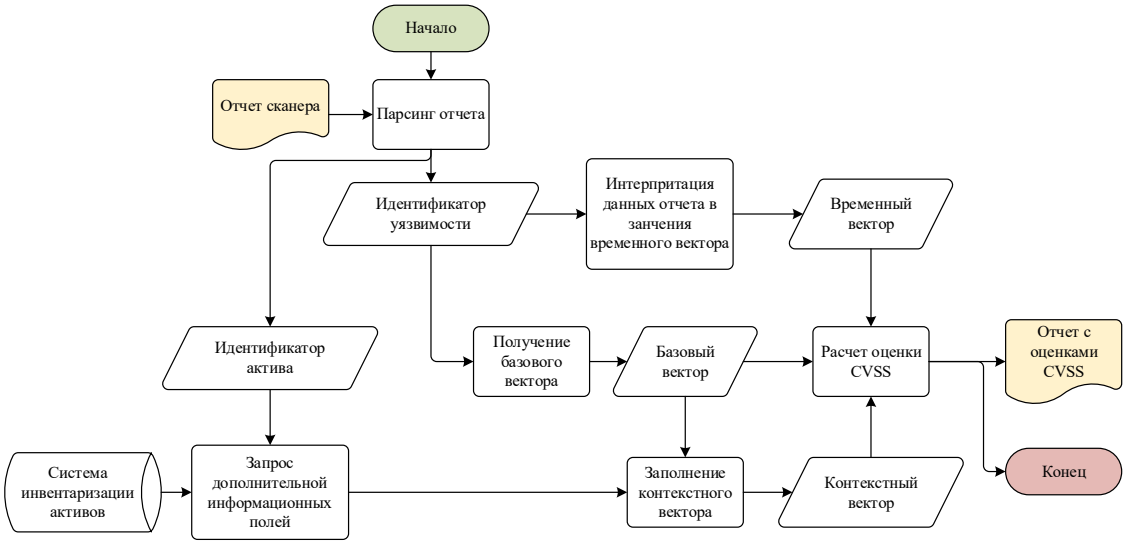


Рисунок 5 – Алгоритм расчета контекстной оценки

В качестве сканера безопасности было выбрано решение Greenbone OpenVAS. Для сравнения в исследование также были включены Max Patrol 8 и RedCheck. Примеры фрагментов структуры отчетов приведены на рисунке 6.

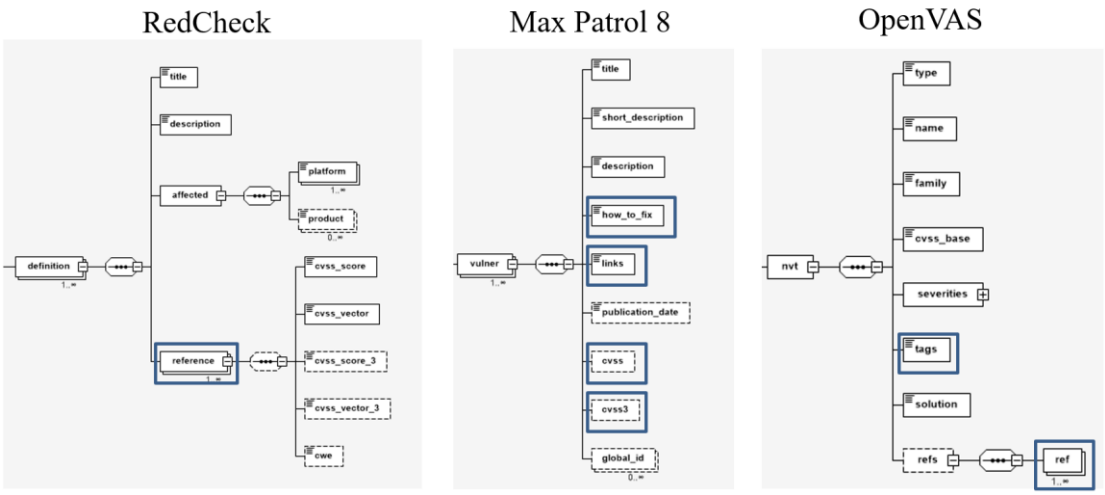


Рисунок 6 – Структуры отчетов сканеров безопасности

На рисунке выделены теги, информацию из которых можно использовать для инициализации метрик временного вектора. Так, например, в OpenVAS каждое значение атрибута type тега solution (Workaround, Mitigation и VendorFix) можно однозначно связать со значениями метрики Remediation Level. Наличие тега ref с значением «cisa» атрибута type (иначе говоря наличие уязвимости в Known Exploited Vulnerability (KEV) catalog) явно указывает на высокий уровень эксплуатируемости рассматриваемой уязвимости, что позволяет явно определить максимальное значение метрики Exploit Code Maturity; также ссылка на бюллетени безопасности производителей программного обеспечения позволяет явно определить значение метрики Report Confidence. Стоит отметить, что некоторые уязвимости в отчетах Max Patrol 8 имеют оценку по временному вектору. RedCheck в рассматриваемом сравнении имеет лишь данные о наличии и типе обновления. На рисунке 7 приведен результат парсинга отчета OpenVAS.

NVT_oid (1)	NVT_name (2)	CVE (3)	KEV (4)	How_to_fix (...	CVSS (6)
1.3.6.1.4.1.25623.1.0.143545	Apache Tomcat AJP RCE Vulnerability (Ghostcat)	CVE-2020-1938	It is in KEV	VendorFix	CVSS:3.1/AV:N/AC:L/PR:N/UI:N/S:U/C:H/I
1.3.6.1.4.1.25623.1.0.103553	DistCC RCE Vulnerability (CVE-2004-2687)	CVE-2004-2687		VendorFix	AV:N/AC:M/Au:N/C:C/I:C/A:C
1.3.6.1.4.1.25623.1.0.106056	VNC Brute Force Login			Mitigation	AV:N/AC:L/Au:N/C:C/I:P/A:P
1.3.6.1.4.1.25623.1.0.103552	PostgreSQL Default Credentials (PostgreSQL Protocol)			Mitigation	AV:N/AC:L/Au:N/C:C/I:P/A:P
1.3.6.1.4.1.25623.1.0.103551	MySQL / MariaDB Default Credentials (MySQL Protocol)			Mitigation	AV:N/AC:L/Au:N/C:C/I:P/A:P
1.3.6.1.4.1.25623.1.0.809883	UnrealIRCd Authentication Spoofing Vulnerability	CVE-2016-7144		VendorFix	CVSS:3.0/AV:N/AC:H/PR:N/UI:N/S:U/C:H/I

Рисунок 7 – Результат парсинга отчета OpenVAS

На рисунке 7 первые три столбца носят исключительно информативный характер и представляют из себя идентификатор уязвимости в нотации OpenVAS, краткое наименование, описывающие суть уязвимости и идентификатор CVE, если таковой имеется. Последние три столбца применяются для автоматизации расчета:

– столбец KEV при наличии в нем значения позволяет заполнить метрику Exploit Code Maturity, также явно говорит о том, что уязвимость подтверждена производителем, что позволяет заполнить метрику Report Confidence;

– столбец How_to_fix однозначно определяет метрику Remediation Level: VendorFix соответствует значению Official Fix, Mitigation – Temporal Fix, Workaround одноименному значению;

– столбец CVSS содержит описательный вектор, который будет расширяться другими метриками в результате автоматизации.

ВЗАИМОДЕЙСТВИЕ С СИСТЕМОЙ ИНВЕНТАРИЗАЦИИ АКТИВОВ

В качестве системы инвентаризации активов была выбрана свободно распространяемая система GLPI. Помимо свободного распространения системы выбор GLPI был обусловлен возможностью расширения функционала за счет плагинов, которые необходимы для введения дополнительных информационных полей.

В рамках исследования для демонстрации работы предложенного подхода активы представляли из себя одноранговую архитектуру. Все дополнительные информационные поля определялись непосредственно в шаблоне карточки актива. Для реализации дополнительных информационных полей использовался плагин Fields [9]. На рисунке 8 приведены включенные в актив дополнительные поля, а также результат запроса к Rest API, по конкретному активу. В дальнейшем полученные результаты интерпретируются в скрипте, отвечающем за расчет полной оценки по CVSS [5–7].

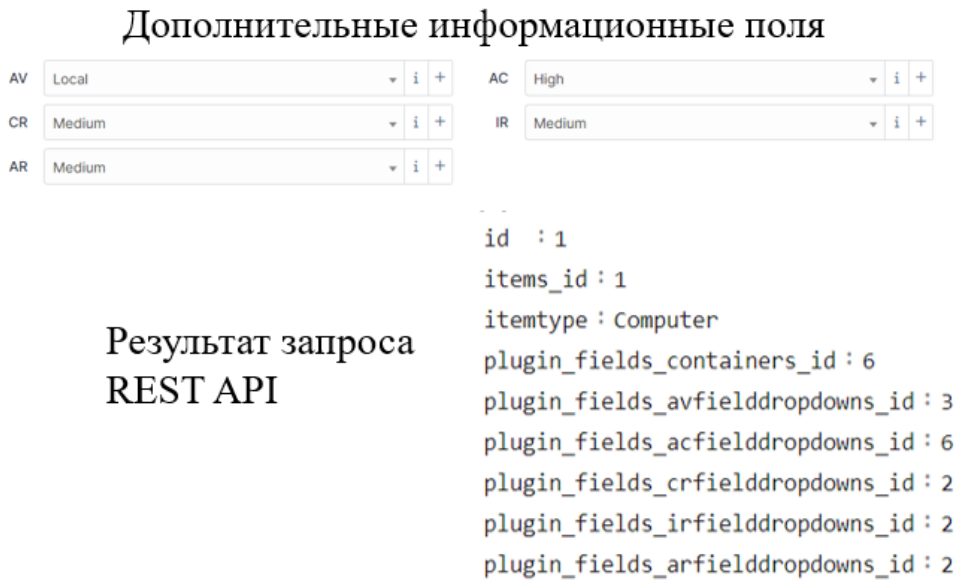


Рисунок 8 – Модернизация системы инвентаризации

Среди метрик контекстного вектора у активов явно определялись метрики требований безопасности (CR, IR, AR), метрика Attack Vector, которая может быть определена в соответствии с положением актива в инфраструктуре, и для изолированных активов, которые помимо средств защиты информации контролируются, например, системой контроля управления доступа, может быть явно задано значение метрики Attack Complexity в значении High. По различным причинам у специалистов может не быть сведений о значении конкретной метрики контекстного вектора, поэтому по умолчанию значения контекстных метрик будут повторять значения метрик базового вектора в соответствии с алгоритмом расчёта, и корректировка оценки будет проводиться за счет критичности актива, определенной ранее.

Упрощенный схема работы скрипта, реализующего алгоритм, определенный ранее, с учетом конкретных инструментов приведен на рисунке 9.

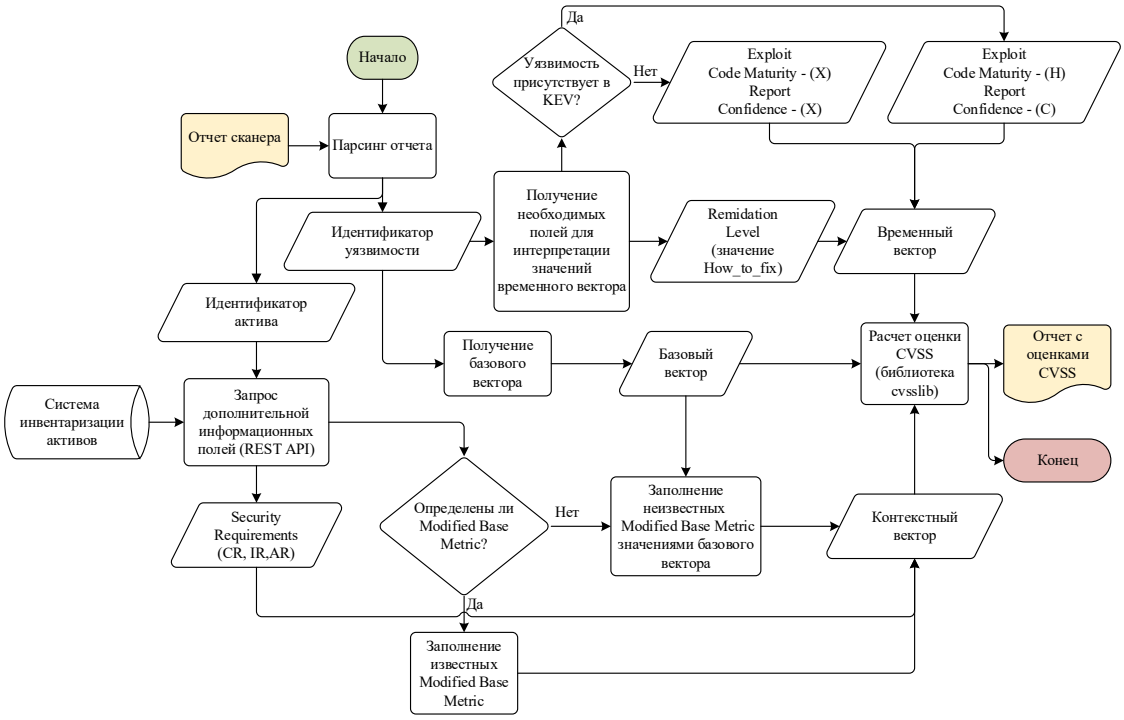


Рисунок 9 – Схема работы скрипта автоматизации

На рисунке 9 в соответствии со спецификацией CVSS (X) соответствует значению Not Defined (X), (H) – High (H), (C) – Confirmed (C). Скрипт реализован на языке программирования Python, за расчет оценки CVSS в котором отвечает библиотека cvsslib [10], в которой реализована логика расчета CVSS.

Результат выполнения расчета полной оценки по CVSS, после выполнения, приведенного выше скрипта представлен на рисунке 10. Расчет выполнялся на примере одного актива, который был зарегистрирован в системе инвентаризации активов

NVT_oid (1)	NVT_name (2)	CVE (3)	Base vector (4)	Temp vecto...	Env vector (6)
1.3.6.1.4.1.25623.1.0.140051	Java RMI Server Insecure Default Configuration RCE Vulnerability	CVE-2011-3556	7.5	5.5	5.2
1.3.6.1.4.1.25623.1.0.901202	The rlogin service is running	CVE-1999-0651	7.5	5.2	4.9
1.3.6.1.4.1.25623.1.0.103482	PHP-CGI-based setups vulnerability when parsing query string	CVE-2012-1823 CVE-2012-2311 CVE-2012-2336 CVE-2012-2335	7.5	5.0	4.7
1.3.6.1.4.1.25623.1.0.140051	Java RMI Server Insecure Default Configuration RCE Vulnerability	CVE-2011-3556	7.5	5.5	5.2
1.3.6.1.4.1.25623.1.0.10498	Test HTTP dangerous methods		7.5	5.2	4.9
1.3.6.1.4.1.25623.1.0.100080	rsh Unencrypted Cleartext Login	CVE-1999-0651	7.5	5.2	4.9
1.3.6.1.4.1.25623.1.0.801281	TWiki Cross-Site Request Forgery Vulnerability - Sep10	CVE-2009-4898	6.8	4.5	4.3

Рисунок 10 – Результаты автоматизации расчета полной оценки CVSS

Первые три столбца носят информативный характер по аналогии с рисунком 7, а столбцы Base vector, Temp vector и Env vector, соответственно, содержат значения оценок CVSS по базовому, временному и контекстному векторам.

Так как каждая версия CVSS содержит схожие метрики, в ходе написания скрипта получилось автоматизировать процесс расчета для версий CVSS от 2 до 3,1.

ПРИМЕНЕНИЕ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ

В рамках проводимого эксперимента использовалась одноранговая инфраструктура, из-за чего инициализировать поля, определяющие метрики контекстного вектора, необходимо на каждом конкретном активе. Для наиболее эффективной инициализации этих метрик необходимо использовать иерархическое представление активов [11], которое основано на положении актива в инфраструктуре. Так, например, если актив находится в изолированном сегменте или под защитой межсетевого экрана, по логике CVSS [5–7] для уязвимости с AV: Network в контекстном векторе MAV по определению принимает значение Adjacent. Соответственно, значения части метрик привязываются не к конкретному активу, а к его положению в иерархии, что в дальнейшем позволит избежать проблем с перемещением активов внутри инфраструктуры и использовать принцип наследования при инициализации значений метрик.

Временной вектор можно дополнить за счет следующих данных:

1) специализированных сервисов, которые предоставляют информацию об эксплуатации уязвимостей в «дикой природе», таких как KEV [12] и AttackerKB [13],

2) агрегаторов данных об уязвимостях, таких как, например, PriON-KB [14], который помимо данных из баз знаний об уязвимостях и социальных сетей предлагает данные о технических отчетах, что является одной из метрик временного вектора.

Однако для взаимодействия через API для некоторых сервисов требуется платная подписка, что может являться существенным ограничением.

СОВЕРШЕНСТВОВАНИЕ ПОДХОДА

Первым и очевидным достоинством полученных оценок перед обычной оценкой по базовому вектору CVSS [5–7] является то, что они отражают опасность уязвимостей, приземленную на целевую инфраструктуру. Даже в таком виде полученные оценки будут являться более объективными для принятия решения о порядке устранения уязвимостей.

Также полученные оценки могут являться входными данными для собственных методик приоретизации [15–20] уязвимостей или для методик, сформированных регуляторами. В качестве методики, предложенной регулятором, можно привести методический документ ФСТЭК «Методика оценки уровня критичности уязвимостей программных, программно-аппаратных средств» [21–22], в котором полная оценка по CVSS является базой для определения критичности и дополнительно корректируется показателем инфраструктуры, введенным методикой. Так как требования по устранению уязвимостей, приведенные в методике, должны выполняться операторами ГИС и КИИ, автоматизация расчетов по CVSS значительно облегчит задачу определения критичности уязвимостей в соответствии с методикой.

Помимо использования полученных оценок в качестве входных данных для различных методик, существует вариант подкрепления данных оценок при помощи дополнительных систем оценивая, например EPSS [23], которая отражает вероятность эксплуатации уязвимости. Схема применения данной системы оценки, совместно с CVSS, приведена на рисунке 11.

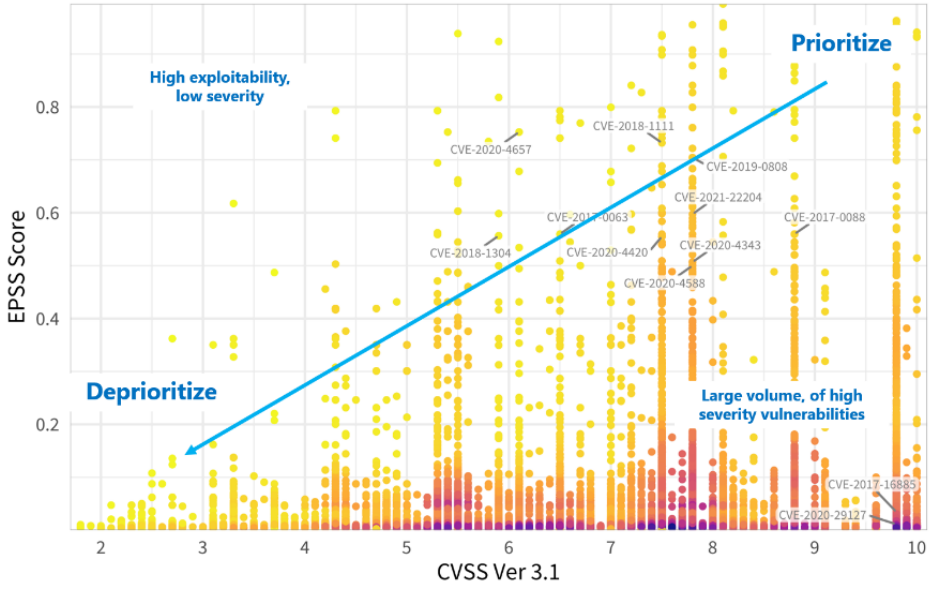


Рисунок 11 – Совместное применение EPSS и CVSS

Так как полученные оценки приземлены на контекст целевой инфраструктуры, применение дополнительных систем оценивания значительно повысит объективность оценки опасности уязвимости.

Стоит отметить, что так как с точки зрения реализации логики автоматизации нет значительных сложностей, то подобную логику можно реализовать в решениях класса SOAR, что позволит сделать процесс оценки полностью автоматическим за счет формирования сценариев.

ЗАКЛЮЧЕНИЕ

Предлагаемый авторами подход к автоматизации представляет из себя быстрое и бесплатное решение задачи определения критичности уязвимостей для их приоритизации и оперативного устранения. Разработанное решение не является полноценной заменой решениям класса Vulnerability Management, но, в сравнении с оценкой только по базовому вектору CVSS, является более объективным за счет приземления на контекст целевой инфраструктуры и частичного учета данных временного вектора.

Основным достоинством предлагаемого подхода к автоматизации расчета полной оценки CVSS является то, что он масштабируемый. Рост уровня зрелости информационной безопасности в организации приводит к росту количества и качества данных необходимых для расчета. Помимо этого, полученные результаты можно использовать как один из критериев критичности уязвимости и расширять его другими в зависимости от уровня зрелости и специфики деятельности организации.

Применение предлагаемого подхода позволит организациям повысить конкурентоспособность организаций за счет оперативного устранения критических уязвимостей, а также может стать отправной точкой для построения эффективного процесса управления уязвимостями.

Список источников

1. Попова, О. Б. Исследование уязвимостей в современном программном обеспечении, используя статистический анализ / О. Б. Попова, Н. В. Кушнир, Ю. С. Носова [и др.] // Современные наукоемкие технологии. – 2021. – № 2. – С. 58–62. – DOI: 10.17513/snt.38494. – EDN FYZKRP.

2. Уязвимости в тренде. – URL: https://www.ptsecurity.com/ru-ru/research/knowledge-base/uyazvimosti-v-trende/ (дата обращения: 01.07.2023).

3. Что такое VPR, и чем он отличается от CVSS. – URL: https://blog.tiger-optics.ru/2021/10/what-is-vpr-and-how-is-it-different-from-cvss/ (дата обращения: 01.07.2023).

4. Qualys TruRisk Platform. – URL: https://www.qualys.com/cloud-platform/ (дата обращения: 01.07.2023).

5. Дойникова, Е. В. Оценка защищенности компьютерных сетей на основе метрик CVSS / Е. В. Дойникова, А. А. Чечулин, И. В. Котенко // Информационно-управляющие системы. – 2017. – № 6 (91). – С. 76–87. DOI: http://dx.doi.org/10.15217/issn1684-8853.2017.6.76. – EDN: ZXWUWH.

6. Дровникова И. Г. Основные виды уязвимостей и взаимосвязь компонентов безопасности при обосновании показателей надёжности системы защиты информации от несанкцио-нированного доступа в автоматизированных системах / И. Г. Дровникова, А. С. Етепнев, Е. А. Рогозин // Приборы и системы. Управление, контроль, диагностика. – 2019. – № 3. – С. 59–64. – DOI: http://dx.doi.org /10.25791/pribor.03.2019.508.

7. Kebande, V. R. CVSS metric-based analysis, classification and assessment of computer network threats and vulnerabilities / V. R. Kebande et al. // 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD). – IEEE, 2018. – P. 1–10.

8. Walkowski, M. Automatic CVSS-based vulnerability prioritization and response with context information / M. Walkowski et al. // 2021 International Conference on Software, Telecommunications and Computer Networks (SoftCOM). – IEEE, 2021. – P. 1–6.

9. Fileds. – URL: https://glpi-plugins.readthedocs.io/en/latest/fields/index.html (дата обращения: 01.07.2023).

10. CVSSlib. – URL: https://pypi.org/project/cvsslib/ (дата обращения: 01.07.2023).

11. Брекоткин, В. Е. Классификация компонентов компьютерной инфраструктуры распределенной организации на основе интеллектуального анализа и структурирования их характеристик / В. Е. Брекоткин, Е. С. Брекоткина, А. С. Павлов, С. В. Павлов // International Journal of Open Information Technologies. – 2022. – Т. 10, № 10. – С. 101–110. – EDN RVACMO.

12. Known Exploited Vulnerabilities Catalog. – URL: https://www.cisa.gov/known-exploited-vulnerabilities-catalog (дата обращения: 01.07.2023).

13. About AttackerKB. – URL: https://attackerkb.com/about (дата обращения: 01.07.2023).

14. PRIOn Knowledge Base. – URL: https://kb.prio-n.com/ (дата обращения: 01.07.2023).

15. Мельников, А. В. Алгоритм оценки относительного уровня опасности совместной эксплуатации уязвимостей информационной безопасности на основе CVSS / А. В. Мельников, В. Е. Чирков // Вестник Воронежского института МВД России. – 2019. – № 1. – С. 37–44. – EDN VXHPCQ.

16. Давлатов, Ш. Р. Анализ защищенности веб-ресурсов на основе метрики CVSS / Ш. Р. Давлатов, П. В. Кучинский // Информатика. – 2020. – Т. 17, № 3. – С. 72–77. – DOI: 10.37661/1816-0301-2020-17-3-72-77. – EDN GFQCXR.

17. Luca, Allodi. Identifying Relevant Information Cues for Vulnerability Assessment Using CVSS / Luca Allodi, Sebastian Banescu, Henning Femmer and Kristian Beckers // Proceedings of the Eighth ACM Conference on Data and Application Security and Privacy (CODASPY '18). Association for Computing Machinery. – New York, NY, USA, 2018. – P. 119–126. https://doi.org/10.1145/3176258.3176340.

18. Dodiya, B. Trend analysis of the CVE classes across CVSS metrics / B. Dodiya, U. K. Singh, V. Gupta // International Journal of Computer Applications. – 2021. – Vol. 975. – P. 8887.

19. Ángel, Jesús Varela-Vaca. Feature models to boost the vulnerability management process / Ángel Jesús Varela-Vaca, Diana Borrego, María Teresa Gómez-López, M. Rafael, A. Gasca, German, Márquez // Journal of Systems and Software. – 2023. – Vol. 195. – P. 111541. https://doi.org/10.1016/j.jss.2022.111541.

20. Зима, В. М. Методика оценивания информационных рисков на основе анализа уязвимостей / В. М. Зима, Р. О. Крюков, А. В. Кравчук // Вопросы оборонной техники. Серия 16: Технические средства противодействия терроризму. – 2019. – № 11–12 (137–138). – С. 36–46. – EDN ZPVNQA.

21. Ефимов, А. О. Концептуальные основы оценки уровня защищенности автоматизированных систем на основе их уязвимости / А. О. Ефимов, И. И. Лившиц, Т. В. Мещерякова, Е. А. Рогозин // Безопасность информационных технологий. – 2023. – Т. 30, № 2. – С. 63–79. – DOI: 10.26583/bit.2023.2.04.

22. Методический документ, методика оценки уровня критичности уязвимостей программных, программно-аппаратных средств. – URL: https://fstec.ru/dokumenty/vse-dokumenty/spetsialnye-normativnye-dokumenty/metodicheskij-dokument-ot-28-oktyabrya-2022-g-2 (дата обращения: 01.07.2023).

23. The EPSS Model. – URL: https://www.first.org/epss/model (дата обращения: 01.07.2023).

References

1. Popova, O. B., Kushnir, N. V., Nosova, Yu. S. et al. Exploring vulnerabilities in modern software using statistical analysis. *Modern High Technologies*, 2021, no. 2, pp. 58–62. DOI: 10.17513/snt.38494. EDN FYZKRP.

2. *Vulnerabilities are trending.* Available at: https://www.ptsecurity.com/ru-ru/research/knowledge-base/uyazvimosti-v-trende/ (accessed 01.07.2023).

3. *What is VPR and how is it different from CVSS.* Available at: https://blog.tiger-optics.ru/2021/10/what-is-vpr-and-how-is-it-different-from-cvss/ (accessed 01.07.2023).

4. Qualys TruRisk Platform. Available at: https://www.qualys.com/cloud-platform/ (accessed 01.07.2023).

5. Doinikova, E. V. Chechulin, A. A., Kotenko, I. V. Computer network security assessment based on CVSS metrics. *Information and control systems*, 2017, no 6 (91), pp. 76–87. DOI: http://dx.doi.org/10.15217/issn1684-8853.2017.6.76. EDN: ZXWUWH.

6. Drovnikova, I. G., Etepnev, A. S., Rogozin, E. A. The main types of vulnerabilities and the relationship of security components in substantiating the reliability indicators of the information protection system from unauthorized access in automated systems. *Devices and systems. Management, control, diagnostics*, 2019, no. 3, pp. 59–64. DOI: http://dx.doi.org/10.25791/pribor.03.2019.508.

7. Kebande, V. R. et al. CVSS metric-based analysis, classification and assessment of computer network threats and vulnerabilities. *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD).* IEEE, 2018, pp. 1–10.

8. Walkowski, M. et al. Automatic CVSS-based vulnerability prioritization and response with context information. *2021 International Conference on Software, Telecommunications and Computer Networks (SoftCOM).* IEEE, 2021, pp. 1–6.

9. *Fileds.* Available at: https://glpi-plugins.readthedocs.io/en/latest/fields/index.html (accessed 01.07.2023).

10. *CVSSlib.* Available at: https://pypi.org/project/cvsslib/ (accessed 01.07.2023).

11. Brekotkin, V. E., Brekotkina, E. S., Pavlov, A. S., Pavlov, S. V. Classification of computer infrastructure components of a distributed organization based on intellectual analysis and structuring of their characteristics. *International Journal of Open Information Technologies*, 2022, vol. 10, no. 10, pp. 101–110. EDN RVACMO.

12. *Known Exploited Vulnerabilities Catalog.* Available at: https://www.cisa.gov/known-exploited-vulnerabilities-catalog (accessed 01.07.2023).

13. *About AttackerKB.* Available at: https://attackerkb.com/about (accessed 01.07.2023).

14. *PROn Knowledge Base.* Available at: https://kb.prio-n.com/ (accessed 01.07.2023).

15. Melnikov, A. V., Chirkov, V. E. Algorithm for assessing the relative level of danger of joint exploitation of information security vulnerabilities based on CVSS. *Bulletin of the Voronezh Institute of the Ministry of Internal Affairs of Russia*, 2019, no 1, pp. 37–44. EDN VXHPCQ.

16. Davlatov, Sh. R., Kuchinskiy, P. V. Analysis of the security of web resources based on the CBSS metric. *Informatiks*, 2020, vol. 17, no. 3, pp. 72–77. DOI: 10.37661/1816-0301-2020-17-3-72-77. EDN GFQCXR.

17. Luca, Allodi, Sebastian, Banescu, Henning, Femmer, and Kristian, Beckers. Identifying Relevant Information Cues for Vulnerability Assessment Using CVSS. *Proceedings of the Eighth ACM Conference on Data and Application Security and Privacy (CODASPY '18). Association for Computing Machinery*. New York, NY, USA, 2018, pp. 119–126. https://doi.org/10.1145/3176258.3176340.

18. Dodiya, B., Singh, U. K., Gupta, V. Trend analysis of the CVE classes across CVSS metrics. *International Journal of Computer Applications*, 2021, vol. 975, p. 8887.

19. Ángel, Jesús Varela-Vaca, Diana, Borrego, María, Teresa Gómez-López, Rafael, M. Gasca, A. German, Márquez. Feature models to boost the vulnerability management process. *Journal of Systems and Software*, 2023, vol. 195, p. 111541. ISSN 0164-1212. https://doi.org/10.1016/j.jss.2022.111541.

20. Zima, V. M., Kriukov, R. O., Kravchuk, A. V. Methodology for assessing information risks based on vulnerability analysis. Voprosy oboronnoi tekhniki. Seriia 16: Tekhnicheskie sredstva protivodeistviia terrorizmu [Issues of Defence Technology. Series 16: Technical means of countering terrorism], 2019, no. 11–12 (137–138), pp. 36–46. EDN ZPVNQA.

21. Efimov, A. O., Livshits, I. I., Meshcheriakova, T. V., Rogozin, E. A. Conceptual framework for assessing the level of security of automated systems based on their vulnerability. Information technology security, 2023, vol. 30, no. 2. pp. 63–79. DOI: 10.26583/bit.2023.2.04.

22. *Methodological document, methodology for assessing the level of criticality of software, firmware and hardware vulnerabilities.* Available at: https://fstec.ru/dokumenty/vse-dokumenty/spetsialnye-normativnye-dokumenty/metodicheskij-dokument-ot-28-oktyabrya-2022-g-2 (accessed 01.07.2023).

23. *The EPSS Model.* Available at: https://www.first.org/epss/model (accessed 01.07.2023).

Статья поступила в редакцию 06.10.2023; одобрена после рецензирования 13.10.2023; принята к публикации 17.10.2023.

The article was submitted 06.10.2023; approved after reviewing 13.10.2023; accepted for publication 17.10.2023.

УДК 004.056.5; 004.052.42

ДОСТОВЕРНОСТЬ КАК СЕРВИС ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ В ЦИФРОВОЙ СРЕДЕ

Ажмухамедов Искандар Маратович, Астраханский государственный университет имени В. Н. Татищева, 414056, Российская Федерация, г. Астрахань, ул. Татищева, 20а,
доктор технических наук, профессор, ORCID: 0000-0001-9058-123X, e-mail: aim_agtu@mail.ru
Хайтул Анастасия Всеволодовна, Астраханский государственный университет имени В. Н. Татищева, 414056, Российская Федерация, г. Астрахань, ул. Татищева, 20а,
бакалавр, ORCID: 0000-0003-2112-8145, e-mail: khaaaytul@icloud.com

В эпоху цифровой информации, где интернет и искусственный интеллект активно используются в обмене данными, их создании и распространении, вопрос обеспечения достоверности информации становится всё более актуальным. Данная работа посвящена проблеме обеспечения достоверности информации в цифровой среде. Сделан акцент на взаимосвязи достоверности с сервисами информационной безопасности: конфиденциальностью, целостностью, доступностью, аутентичностью и неотказуемостью. Новизна работы заключается в комплексном исследовании понятия «достоверность» и выдвижении идеи рассмотрения его как сервиса информационной безопасности. Рассмотрены различные определения понятия «достоверность» и сформулировано собственное определение, подчеркивающее инвариантные особенности данного понятия. В работе также рассмотрены методы манипуляции сознанием, широкий спектр методов искажения информации и использование искусственного интеллекта как инструмента для намеренного создания убедительных фейков. Остро подчеркнута необходимость разработки методик и технологий, обеспечивающих проверку достоверности информации в цифровой среде, особенно учитывая всё более широкое применение искусственного интеллекта для генерации и распространения фейков. Намечены потенциальные пути интеграции современных технологий, таких как блокчейн, в процессы защиты данных от искажений и несанкционированных изменений для обеспечения их целостности и достоверности.

Ключевые слова: информационная безопасность, достоверность информации, деструктивное влияние, свойства информационной безопасности, цифровая среда

RELIABILITY AS AN INFORMATION SECURITY SERVICE IN THE DIGITAL ENVIRONMENT

Azhmukhamedov Iskandar M., Astrakhan Tatishchev State University, 20a Tatishchev St., Astrakhan, 414056, Russian Federation,
Doct. Sci. (Engineering), Professor, ORCID: 0000-0001-9058-123X, e-mail: aim_agtu@mail.ru
Khaytul Anastasia V., Astrakhan Tatishchev State University, 20a Tatishchev St., Astrakhan, 414056, Russian Federation,
bachelor, ORCID: 0000-0003-2112-8145, e-mail: khaaaytul@icloud.com

In the era of digital information, where the Internet and artificial intelligence are actively used in data exchange, creation and dissemination, the issue of ensuring the reliability of information is becoming more and more relevant. This paper is devoted to the problem of information assurance in the digital environment. It emphasizes the relationship between trustworthiness and information security services: confidentiality, integrity, availability, authenticity and unreliability. The novelty of the work lies in the comprehensive study of the concept of "trustworthiness" and the idea of considering it as an information security service. Various definitions of the concept of "trustworthiness" are considered and our own definition is formulated, emphasizing the invariant features of this concept. The paper also considers methods of consciousness manipulation, a wide range of methods of information distortion and the use of artificial intelligence as a tool for deliberately creating convincing fakes. The need to develop methods and technologies to ensure the verification of information in the digital environment, especially given the increasing use of artificial intelligence for the generation and dissemination of fakes, is acutely emphasized. Potential ways to integrate modern technologies, such as blockchain, into the processes of protecting data from distortion and unauthorized changes to ensure its integrity and trustworthiness are outlined.

Keywords: information security, information reliability, destructive influence, information security properties, digital environment

ВВЕДЕНИЕ

Обмен информации является важнейшей частью существования человеческого общества. Интенсивность такого обмена растет в том числе из-за бурного внедрения в жизнь человека информационных технологий, в особенности интернета. Как показывает исследование креативного агентства «We Are Social» и сервиса для SMM «Meltwater» «Digital 2023 Global Overview Report», 64,4 % из 8,01 млрд жителей нашей планеты, т. е. 5,16 млрд человек. используют интернет. Согласно

этим же исследованиям, на начало 2023 г. социальными сетями пользовались около 4,76 млрд человек [1]. Бурное развитие информационных технологий, в свою очередь, сделало весьма актуальными вопросы, связанные с обеспечением информационной безопасности (ИБ).

При классическом подходе под информационной безопасностью обычно понимают такие сервисы, как конфиденциальность, целостность, доступность. Для определенного круга задач к ним добавляются аутентичность и неотказуемость. Таким образом, в сферу интересов ИБ входят сервисы: конфиденциальность, целостность, доступность, аутентичность и неотказуемость. Однако все они относятся к проблеме сохранения собственной информации. В то же время в последние годы все более актуальной становится проблема проникновения в общественную среду «внешней», часто оказывающей деструктивное влияние информации. Распространение подобной информации способствует увеличению энтропии (степени хаоса) в социальной системе. Также циркуляция недостоверной информации приводит к лавинообразному процессу искажения «истины» и, как следствие, неадекватному поведению, часто носящему весьма разрушительный, непредсказуемый характер (уличные беспорядки, массовое насилие и т. п.).

Как указывал в своей работе «Манипуляция сознанием» С. Г. Кара-Мурза, «человек не просто социальное существо, которое может существовать, только интенсивно обмениваясь информацией с себе подобными … Он обладает разумом, способным к мышлению, речью и языком. Язык и мышление – большие сложные системы, на которые можно воздействовать с целью программирования поведения человека». По мнению автора, влиять на поведение человека можно различными способами, например, при помощи воздействия на его биологические процессы и структуры, но данные способы являются преступным вмешательством в организм человека [2]. Поэтому для воздействия на сознание часто используют информацию, которая может оказывать деструктивное влияние.

Наиболее ярко деструктивные воздействия проявляются во время ведения «информационных войн». Проблему данного термина впервые начал изучать американский психолог и социолог Гарольд Ласуэлл ещё во времена Первой мировой войны. Г. Ласуэлл считал, что пропаганда является одним из самых «действенных орудий». В своих работах он рассматривал различные методы пропаганды, которая поднимает моральный дух собственного народа и понижает его у противника. Это, в свою очередь, оказывает решающее воздействие и обеспечивает достижение поставленных государством целей [3].

Таким образом, информационная манипуляция является действенным оружием и может использоваться в разрушительных целях, во вред обществу. Поэтому защита населения от подобного воздействия является актуальной задачей.

Исходя из этого, на федеральном уровне в Российской Федерации был принят ряд законов, регламентирующих распространение «чувствительной» информации. Были внесены многочисленные поправки в Федеральный закон «Об информации, информационных технологиях и о защите информации» от 27.07.2006 № 149-ФЗ. Например, внедрение так называемого закона И. Яровой, обязывающего операторов связи и интернет-компаний хранить текстовые сообщения, голосовую информацию, изображения, звуки, видео и иные электронные сообщения пользователей в срок до шести месяцев [4] или введение «Единого реестра доменных имен, указателей страниц сайтов в сети «Интернет» и сетевых адресов, позволяющих идентифицировать сайты в сети «Интернет», содержащие информацию, распространение которой в Российской Федерации запрещено» [5] и т. п.

При этом необходимо отметить, что существенная часть деструктивной информации является по своей сути недостоверной (искажает данные, факты, сведения, географические координаты и время имевших место событий).

Исходя из вышеизложенного, представляется целесообразным расширить перечень свойств информации, рассматриваемых в рамках ИБ, включив в него понятие «достоверность», поскольку оно является одним из важнейших имманентных (внутренне присущих объекту) свойств информации и может оказывать существенное влияние на безопасность как отдельных пользователей, так и общества в целом.

ОПРЕДЕЛЕНИЕ ПОНЯТИЯ «ДОСТОВЕРНОСТЬ»

В толковых словарях и энциклопедиях [6–8] приводятся различные определения понятия «достоверность».

Так, например, в энциклопедическом словаре [6] говорится, что «достоверность – это форма существования истины, обоснованной каким-либо способом (напр., экспериментом, логическим доказательством)» [9].

В словаре логики [7] дается следующее определение: «Достоверность – обоснованность, доказательность, бесспорность знания» [10].

В философском энциклопедическом словаре [8] достоверность определяется как: «убеждение, основанное на знании и исключающее всякое сомнение. Достоверность может быть субъективной

(в вере), объективной (в науке), непосредственной (основанной на созерцании, собственном восприятии, собственном переживании – интуитивная достоверность) или опосредованной, исторической или логической (полученной посредством сообщения или посредством мышления)» [11].

Также широко известен толковый социологический словарь [12], в котором сказано, что «достоверность (reliability) – надежность собранных данных либо испытания или оценки, их сбора» [13].

Кроме этого, существует множество других определений термина достоверность. Но несмотря на такое разнообразие, во всех вышеописанных определениях имеются общие, инвариантные черты, которые подчеркивают степень надежности, правдоподобия информации, знаний или данных. Помимо этого, во всех определениях подчеркивается, что достоверность связана с уверенностью в том, что информация или знания являются точными и могут быть доказаны, исключая сомнения в их ненадежности.

Исходя из всего вышеизложенного, предлагается следующее определение достоверности как свойства безопасности информации, препятствующего деструктивному воздействию на потребителей данной информации: достоверной считается информация, точно и объективно отражающая события, произошедшие в конкретном (соответствующем действительности) месте и в указанное время.

Согласно данному определению, важно, чтобы сведения о событии, месте и времени были достоверными в совокупности. Например, распространяемая информация может являться полностью ложной (выдуманной), либо событие могло действительно произойти, но не в том месте и/или не в то время. В этом случае распространяемая информация может считаться лишь частично достоверной. Степень доверия к такой информации и, как следствие, к её источнику будут подорваны. Если такая ситуация возникла в результате непреднамеренной ошибки в освещении того или иного события, необходимо оперативно её устранить и довести до сведения пользователей причину её возникновения, а также предпринять комплекс мер для недопущения в дальнейшем подобных ошибок.

ИСКАЖЕНИЕ СУТИ СОБЫТИЯ ИЛИ ФАКТА

Отдельного рассмотрения требует случай искажения сути произошедшего события. Природа данного искажения может носить как объективный, так и субъективный характер. Объективное (непреднамеренное) искажение обычно возникает на начальной стадии освещения того или иного события или факта, когда высока степень неопределенности, неизвестны многие факторы, характеризующие произошедшее.

Кроме этого, может иметь место субъективное (преднамеренное) искажение фактов. Источник, распространяющий информацию в интернет, может сознательно вводить в заблуждение пользователей различными способами, например:

1. Выборочное представление фактов. Автор представляет только часть информации, исключая или добавляя те сведения, которые могли бы изменить общее восприятие ситуации. Такая подача информации приводит к искажённому представлению о происходящем, а также подрывает доверие к источнику информации.

Один из примеров подобного представления информации – это ситуация, произошедшая с NBC Nightly News и Брайаном Уильямсом.

В 2015 г. Брайан Уильямс, ведущий NBC Nightly News, был изобличен в ложных утверждениях о своем участии в событиях в Ираке. Он утверждал, что его вертолет был подвергнут обстрелу противником, что оказалось неправдой. Этот скандал стал поводом для дальнейшего расследования и привел к выявлению других случаев, когда Уильямс представлял события в искаженном свете.

Этот случай широко обсуждался в СМИ и соцсетях и подорвал доверие к NBC Nightly News как к источнику новостей. Брайан Уильямс временно был отстранен от работы, о чем NBC Nightly News оповестил пользователей [14, 15].

2. Использование эмоциональных аргументов. Автор усиливает эмоциональные аспекты событий, чтобы вызвать определенные реакции у читателей и увести их от объективного видения ситуации.

3. Интерпретация событий. Вместо объективной картины произошедшего автор в действительности излагает свой собственный взгляд и дает собственную (не всегда легко обнаруживаемую) интерпретацию событий. В этом случае мы имеем дело с так называемыми оценочными суждениями, которые являются «симбиозом» фактов и субъективного мнения о них, опирающегося на личный опыт человека. При этом признаками оценочных суждений часто являются:

– морализаторство, т. е. навязывание какого-либо мнения при помощи использования устоявшихся моральных ценностей;

– категоричность, отличительной чертой которой является использование слов «должен», «обязан» и т. п.;

– некорректное обобщение, т. е. распространение каких-либо выводов на генеральную совокупность на основе отдельных исключительных событий или фактов.

Часто бывает так, что оценочное суждение, агрессивно подкреплённое автором ссылкой на свой «большой» опыт в дальнейшем, воспринимается как достоверный факт.

Таким образом, для увеличения степени достоверности информации необходимо при её передаче обеспечить минимальную субъективность. По крайней мере, нужно четко различать факты и их интерпретацию. Высказывание типа «стакан на половину полон» или «на половину пуст» является примером интерпретации. В данном случае корректнее было бы сказать, что в стакан налито определенное количество жидкости.

Ярким примером неверной интерпретации информации из реальной жизни является случай, когда сотрудники грузинского телеканала «Имеди» записали сюжет с возможными сценариями развития событий. Один из таких сценариев допускал начало российского вторжения в Грузию, а также гибель президента страны Михаила Саакашвили. Несмотря на то, что данный репортаж являлся лишь размышлениями о возможных вариантах развития событий, зрители, подключившиеся к трансляции не с самого её начала, интерпретировали увиденные кадры как реально происходящие события, что вызвало панику в стране [16].

4. Некорректное цитирование источников и использование медиаматериалов. Автор некорректно ссылается на авторитетные источники, пытаясь усилить степень доверия к передаваемой информации. С этой же целью часто используются различные медиаматериалы (изображения, видеоролики или аудиозаписи), которые могут не иметь отношение к передаваемой автором информации или могут быть им целенаправленно искажены, чтобы поддержать определенную точку зрения на описываемые события или факты.

Все вышеуказанные методы могут использоваться как в отдельности, так и в совокупности для создания иллюзии достоверности, в то время как сама информация может быть ложной или искаженной.

НЕДОСТОВЕРНОЕ УКАЗАНИЕ МЕСТА

Недостоверное указание места, где произошло то или иное событие, подрывает доверие к источнику, особенно если эта информация является важной для понимания контекста события. Например, в 2009 г. CNN опубликовала фотографии, которые, как утверждалось, были сделаны во время бомбардировки в Газе [17]. Однако позже выяснилось, что одна из фотографий была сделана не в Газе, а в Ливане в 2006 г. Позже данные заменили, но подобные ошибки в указании местоположения часто вызывают недовольство и недоверие у читателей в дальнейшем.

НЕДОСТОВЕРНОЕ УКАЗАНИЕ ВРЕМЕНИ

Доверие к информации может быть подорвано также некорректным указанием времени, когда произошло то или иное событие. Такие ошибки в указании времени могут возникать в результате влияния человеческого фактора, технических сбоев, задержек при получении сведений и т.п.

Например, первоапрельская шутка, опубликованная позже или на пару дней раньше, может быть воспринята большинством пользователей цифровых сервисов как реально случившееся событие. Одним из ярких примеров подобного события является опубликованная 31 марта новость о возобновлении полетов самолета «Конкорд». Этой неожиданной и кажущейся правдоподобной новостью, удалось запутать агентство France-Presse. Лайнер должен был совершить свой первый полет с 2003 года в салоне «Ле-Бурже» 16 июня 2009 года. Однако это событие оказалось лишь первоапрельским розыгрышем, придуманным парижским музеем авиации и космонавтики. Музей, как выяснилось, стремится вернуть «Конкорд» в небо, но пока не обладает необходимыми ресурсами для этой цели [18].

К этой же категории неполностью достоверной информации относится использование так называемых «консервов», когда распространяется информация (обычно в формате видеофайла), которая была заранее отснята, но выпускается в эфир с указанием нужной даты.

Кроме того, следует отметить, что достоверность может быть как объективной, так и субъективной. Объективная достоверность так же, как и абсолютная истина, недостижима. Как указывал еще Ф. Ницше: «Фактов не существует, есть только их интерпретации». Люди принимают решения, основываясь на собственной (субъективной) уверенности в достоверности полученной информации. То есть их поступки основаны на том, на сколько информация кажется им надежной. На это, в свою очередь, оказывают сильное влияние личные убеждения, опыт и предпочтения человека. При этом одно и то же утверждение, которое кажется одному человеку достоверным, другому может показаться сомнительным.

Субъективная оценка степени достоверности информации зависит от ряда факторов:

1. Предвзятость. Люди могут иметь предвзятое мнение, которое влияет на их восприятие информации. Например, если у человека есть политические или религиозные предубеждения, он более склонен относиться к информации в соответствии со своими представлениями.

2. Эмоции. Эмоциональное состояние также оказывает влияние на восприятие и оценку информации. Яркие эмоции, такие как страх, гнев или энтузиазм, сильно мешают объективно анализировать информацию.

3. Личный опыт. Личный опыт каждого человека уникален. Люди в гораздо большей степени доверяют информации, которая соответствует их собственному опыту, и сомневаются в информации, которая ему противоречит.

4. Социальное окружение. Мнения, распространённые в окружающей социальной среде, также оказывают сильное воздействие на процесс восприятия информации. Обычно человек склонен больше доверять информации, которая соответствует мнению большинства в его социальной группе.

5. Индивидуальные ценности и убеждения. Информация, которая соответствует этическим или моральным убеждениям человека, обычно воспринимается им как более достоверная.

6. Информационные источники. Люди склонны переносить доверие к источникам информации на саму информацию, т. е. если данные получены из надежного, по их мнению, источника, то она скорее всего будет воспринята как достоверная.

7. Форма представления. Информация, представленная в форме научного исследования, в большинстве случаев считается более достоверной, чем информация, представленная в виде частного мнения, изложенного в социальных сетях.

Таким образом, можно утверждать, что степень достоверности зависит от следующего множества факторов:

$$< B; F; E; P; C; S; R >, \quad (1)$$

где B – Bias (предвзятость);

F – Feelings (эмоции);

E – Experience (опыт);

P – Perception (восприятие);

C – Conviction (убеждение);

S – Source (источники);

R – Representation (форма представления).

Таким образом, информация, циркулирующая в цифровой среде, может быть как полностью, так и частично недостоверной, что с учетом субъективных особенностей пользователей цифровых сервисов часто приводит к принятию ими неверных решений и может оказывать на них деструктивное воздействие. Поскольку довольно часто распространение такого рода информации носит вирусный (лавинообразный) характер, необходимо на ранних стадиях оперативно выявлять и принимать контрмеры по их нейтрализации. И несмотря на то, что имеется ряд работ, посвящённых созданию автоматических/автоматизированных систем распознавания экстремистской и иной деструктивной информации [19–22], вопросы анализа достоверности публикуемых в сети данных на сегодняшний день не нашли должного отражения. Таким образом, возникает задача создания методики и соответствующего ПО для пресечения вирусного распространения недостоверной информации среди пользователей цифровых сервисов.

ВЗАИМОСВЯЗЬ ДОСТОВЕРНОСТИ И ДРУГИХ СЕРВИСОВ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

Рассмотрим взаимосвязь достоверности с основными сервисами информационной безопасности, которые были перечислены ранее: конфиденциальность, целостность, доступность, аутентичность и неотказуемость.

Конфиденциальность как сервис информационной безопасности представляет собой такое свойство информации, которое обеспечивает недоступность сведений для нелегитимного пользователя.

Требования к достоверности и конфиденциальности часто противоречат друг другу, поскольку открытое, неограниченное распространение информации в значительной мере способствует выявлению факта ее недостоверности. В случае с конфиденциальной информацией такого рода возможности весьма ограничены.

Целостность является свойством информации, означающим, что данные остаются неискажёнными в течение всего их жизненного цикла, т. е. они не подверглись несанкционированным изменениям, искажениям или посторонним вмешательствам.

Необходимо отметить, что лишь целостность и достоверность являются имманентными свойствами информации. Все остальные сервисы, указанные выше, обеспечивают защиту информации при помощи внешних механизмов, задействованных лицом, принимающим решение. Для защиты информации может быть использовано хэширование, шифрование, ЭЦП, резервирование каналов приёма-передачи и т. д.

Взаимосвязь между целостностью и достоверностью заключается в том, что нарушение целостности ведет к тому, что информация часто становится недостоверной.

Доступность – это свойство информации, которое обеспечивает гарантию получения данных легитимным пользователям в любое время.

Потеря доступности к объективной информации способствует увеличению спроса на альтернативные информационные ресурсы. Однако они могут быть менее проверенными и информация,

полученная из такого рода источников, часто может оказаться недостоверной, что впоследствии повышает вероятность появления и распространения такой информации в цифровой среде. Следовательно, обеспечение доступности к объективной информации играет ключевую роль в сдерживании распространения фейковых (недостоверных) сведений.

Кроме этого, доступность помогает пользователям в проверке достоверности данных. То есть при открытом доступе к информации люди могут свободно убедиться в том, что событие реально случилось.

Аутентичность и неотказуемость – взаимодополняющие свойства, которые направлены на обеспечение доверия к происхождению и целостности информации. Аутентичность помогает убедиться в идентификации субъекта, распространяющего информацию, и проверяет не было ли подделано или изменено авторство в процессе распространения данных. Неотказуемость обеспечивает невозможность отказа от передачи или приема информации субъектами информационных отношений.

Нарушение этих сервисов так же, как и в случае нарушения сервиса «доступность», создает благоприятные условия и провоцирует распространение недостоверной информации, поскольку у создателей такого рода информации возникает чувство безнаказанности (в большинстве своем вполне обоснованное).

Взаимосвязь достоверности и других свойств ИБ можно представить в виде схемы, приведенной на рисунке.

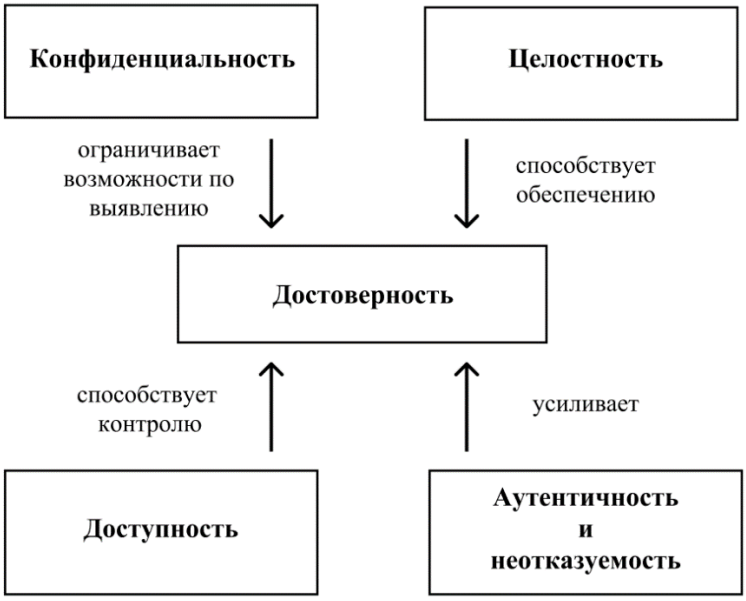


Рисунок – Взаимосвязь достоверности и классических сервисов информационной безопасности

Таким образом, существует взаимосвязь между достоверностью и классическими сервисами ИБ. Нарушения последних приводят к распространению ложной (недостоверной) информации, что часто влечет негативные последствия.

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ЕГО ВЛИЯНИЕ НА ДОСТОВЕРНОСТЬ ИНФОРМАЦИИ

Важно отметить, что в цифровой среде начинают пользоваться большим спросом средства, использующие искусственный интеллект (ИИ). Он является одной из самых перспективных и быстроразвивающихся технологией. Концепция искусственного интеллекта была впервые озвучена в 1950-х гг. Аланом Тьюрингом [23], а первые идеи о попытке проведения исследований с ИИ были предложены в 1956 г. во время конференции в Дартмуте [24]. С тех пор разработка искусственного интеллекта ушла далеко вперед, и сегодня ИИ имеет многочисленные применения в различных отраслях, таких как здравоохранение, финансы, образование и т. д.

Среди технологий искусственного интеллекта особое внимание уделяется машинному обучению (Machine learning) и глубокому обучению (Deep learning). Машинное обучение позволяет компьютерам учиться на основе данных, маркированных человеком, а глубокое обучение использует алгоритмы, имитирующие нейронные связи человеческого мозга.

Ниже представлены примеры искусственного интеллекта:

– OpenAI GPT-3: технология, способная генерировать тексты, имитируя человеческий язык. Обычно она используется для создания контента, перевода текста, ответов на вопросы и т. п.

– WaveGAN: нейронная сеть, способная генерировать аудио. WaveGAN может использоваться для создания музыкальных треков, звуковых эффектов и других аудиозаписей.

– DeepArt.io: сервис, который использует нейронные сети для создания художественных изображений на основе пользовательских фотографий.

В эпоху цифровизации ИИ оказывает всё большее влияние на мир информации, способствуя созданию, анализу, фильтрации и распространению данных с высокой скоростью. Однако вместе с новыми возможностями появляются и новые риски, связанные с достоверностью информации.

Современные технологии ИИ, такие как генеративно-состязательные сети (GANs), способны создавать гиперреалистичные изображения, тексты и видео, что может приводить к распространению ложной или искажённой информации.

Так, например, DeepFakes использует ИИ и нейронные сети для создания реалистичных, но сфабрикованных видео. Принцип работы заключается в том, что пользователь загружает фото человека, с которым хочет сгенерировать видео, и нейросеть при помощи алгоритмов заменяет лица людей и формирует фальшивый ролик.

Другим примером использования искусственного интеллекта для создания заведомо ложной информации является нейросеть Text-to-Speech. С ее помощью человек создает синтетическую речь, которая сформировывает фейковые аудиофайлы.

Кроме генерации видео- и аудиофайлов, нейросети могут создавать изображения по текстовым запросам пользователя, тем самым увеличивая риск появления не соответствующей действительности информации. Одной из таких популярных нейросетей является Midjourney. Подобные технологии могут быть использованы для создания дезинформации и «фейковых новостей», что, в свою очередь, может сказаться на общественном мнении или же повлиять на процессы, связанные с различными сферами жизнедеятельности человека.

В мире уже были зафиксированы случае, где искусственный интеллект был использован для создания и распространения фальшивых новостей. Так, в Китае впервые задержали человека за распространение фейковой новости, сгенерированной при помощи чат-бота ChatGPT [25].

Важно понимать, что алгоритмы машинного обучения анализируют огромные объемы данных для выявления закономерностей и зависимостей, на основании которых формируют предсказания или принимают решения. Однако качество этих предсказаний и решений напрямую зависит от качества анализируемых данных и корректности обучения модели. Если данные, на которых обучается модель, искажены, неполны или содержат ошибки, то модель может научиться неправильным зависимостям и закономерностям, что ведет к неверным или искаженным выводам. Это явление известно как «мусор на входе – мусор на выходе» (garbage in, garbage out) [26].

Также стоит отметить, что многие алгоритмы машинного обучения, особенно в области глубокого обучения, работают как «чёрные ящики», т. е. даже разработчики в большинстве случаев не понимают, как и почему модель пришла к тому или иному выводу. Это создает проблемы в интерпретации результатов и проверке достоверности предсказаний модели.

Всё вышенаписанное подчеркивает важность использования качественных данных для обучения модели и разработки методов интерпретации результатов машинного обучения при применении технологий ИИ, чтобы минимизировать риск распространения недостоверной информации.

Для борьбы с подобными проблемами важно разрабатывать и внедрять методы и технологии, способные обеспечивать проверку, подтверждение достоверности информации, а также выявление искаженных или сфабрикованных данных.

Использование блокчейн-технологий могло бы обеспечивать целостность и неизменность данных, что особенно важно для обеспечения достоверности информации в цифровой среде. Блокчейн – это технология, которая позволяет хранить данные в виде последовательности блоков, каждый из которых содержит информацию и ссылку на предыдущий блок, формируя тем самым цепочку. Такая технология оказывается весьма перспективной для создания систем, обеспечивающих защиту данных от несанкционированных изменений и подделок.

Хотя современные технологии, включая блокчейн, предоставляют нам новые возможности для создания прозрачных и надежных систем, где достоверность информации может быть эффективно проверена и подтверждена, это все равно порождает новые риски. В условиях, когда искусственный интеллект и другие передовые технологии проникают во все сферы нашей жизни, важность вопросов, связанных с подтверждением подлинности и достоверности информации, усиливается.

ЗАКЛЮЧЕНИЕ

Обеспечение достоверности информации в цифровую эпоху становится все более актуальной задачей, объединяющей в себе различные аспекты в области информационной безопасности. Техники манипулирования сознанием и методы противодействия им охватывают широкий спектр

методов, начиная от управления эмоциональным фоном и заканчивая использованием современных технологий ИИ для создания убедительных фейков.

Тесная взаимосвязь достоверности с другими сервисами информационной безопасности подчеркивает сложность и многофакторность задачи ее обеспечения в цифровом пространстве.

При этом искусственный интеллект выступает в двух различных (противоположных) ипостасях: с одной стороны, он предоставляет злоумышленникам широкие возможности для создания весьма реалистичных фейков (например, такие технологии, как GANs и DeepFakes); с другой – технологии ИИ могут быть использованы для идентификации недостоверной информации.

Однако, несмотря на все технологические инновации, человеческий фактор, этика и ответственность остаются основополагающими в вопросе обеспечения достоверности информации. Усилия по обеспечению достоверности данных в цифровой среде должны опираться как на технические и технологические аспекты, так и на развитие культуры цифровой грамотности и осмысленного подхода в использовании и распространении информации в цифровой среде.

Список источников

1. DIGITAL 2023: GLOBAL OVERVIEW REPORT // datareportal. – URL: https://datareportal.com/reports/digital-2023-global-overview-report (дата обращения: 08.08.2023).

2. Кара-Мурза, С. Г. Манипуляция сознанием / С. Г. Кара-Мурза. – 2003. – 464 с. – URL: http://lib.ru /POLITOLOG/karamurza.txt (дата обращения: 10.08.2023).

3. Красовская, Н. Р. К вопросу классификации информационных войн / Н. Р. Красовская, А. А. Гуляев // Социология науки и технологий. – 2019. – Т. 10, № 2. – С. 44–55. – DOI: 10.24411/2079-0910-2019-12002.

4. Статья 10.1. Обязанности организатора распространения информации в сети «Интернет». – URL: https://www.consultant.ru/document/cons_doc_LAW_61798/9ab7abe2b9fe407f507610f7e6e14a951d575585/ (дата обращения: 13.08.2023).

5. Статья 15.1. Единый реестр доменных имен, указателей страниц сайтов в сети «Интернет» и сетевых адресов, позволяющих идентифицировать сайты в сети «Интернет», содержащие информацию, распространение которой в Российской Федерации запрещено. – URL: https://www.consultant.ru/document/cons_doc_LAW_61798/38c8ea666d27d9dc12b078c556e316e90248f551/#dst16 (дата обращения: 13.08.2023).

6. Большой энциклопедический словарь // Wikipedia. – URL: https://ru.wikipedia.org/wiki/Большой_энциклопедический_словарь (дата обращения: 21.08.2023).

7. Ивин, А. А. Словарь по логике / А. А. Ивин, А. Л. Никифоров. – Москва : Гуманитарный издательский центр ВЛАДОС, 1997. – 384 с.

8. Философский энциклопедический словарь // Словари онлайн. – URL: https://rus-philosophy-enc.slovaronline.com/ (дата обращения: 21.08.2023).

9. Сборник энциклопедий и словарей русского языка // Значение слова Достоверность в энциклопедическом словаре. – URL: https://diclist.ru/slovar/enciklopedicheskiy/d/1-dostovernost.html (дата обращения: 21.08.2023).

10. Сборник энциклопедий и словарей русского языка // Значение слова Достоверность в словаре логики. – URL: https://diclist.ru/slovar/logiki/d/dostovernost.html (дата обращения: 21.08.2023).

11. Сборник энциклопедий и словарей русского языка // Значение слова Достоверность в философском словаре. – URL: https://diclist.ru/slovar/logiki/d/dostovernost.html (дата обращения: 21.08.2023).

12. Большой толковый социологический словарь // gufo.me. – URL: https://gufo.me/dict /social_dict?page=5&letter=д (дата обращения: 21.08.2023).

13. Большой толковый социологический словарь // Словари онлайн. – URL: https://rus-philosophy-enc.slovaronline.com/ (дата обращения: 21.08.2023).

14. NBC suspends journalist Brian Williams for 'lying' about brush with death in Iraq war // SCMP. – URL: https://www.scmp.com/news/world/article/1709945/nbc-suspends-journalist-brian-williams-lying-about-brush-death-iraq-war (дата обращения: 19.09.2023).

15. Телеведущий телеканала NBC извинился за ложь в эфире и прервал работу // РИА новости. – URL: https://ria.ru/20150208/1046536663.html?ysclid=ln0b9d9wgm318468016 (дата обращения: 19.09.2023).

16. Как телеканал «Имеди» президента Грузии «убил» // Lenta.ru: сайт. – URL: https://lenta.ru/articles/2010/03/15/imedi/] (дата обращения: 20.09.2023).

17. Israel bombs as tanks wait on outskirts of Gaza // CNN. – URL: https://edition.cnn.com/2009/WORLD/meast/01/02/israel.gaza/index.html (дата обращения: 20.09.2023).

18. Французский авиамузей разыграл журналистов, пообещав полет «Конкорда» // РИА новости. – URL: https://ria.ru/20090401/166698857.html?ysclid=lmz0yz41n5231955908 (дата обращения: 21.09.2023).

19. Свидетельство о государственной регистрации программы для ЭВМ № 2021680437 Российская Федерация. Программа для парсинга сообщений в публичных телеграм-каналах. – № 2021669987 ; заявл. 30.11.2021 ; опубл. 10.12.2021 / И. И. Карабак, К. А. Зорин, И. М. Ажмухамедов.

20. Минаев, В. А. Автоматизированное выявление деструктивного контента в социальных медиа / В. А. Минаев, А. В. Симонов, А. Д. Реброва // Информационная безопасность: вчера, сегодня, завтра : сборник статей по материалам IV Международной научно-практической конференции, Москва, 22 апреля 2021 года / под ред. В. В. Арутюнова. – Москва : Российский государственный гуманитарный университет, 2021. – С. 124–130.

21. Вепрев, С. Б. Применение технологии Big Data для выявления деструктивного контента в Интернете / С. Б. Вепрев, С. А. Нестерович // Информационная безопасность: вчера, сегодня, завтра : сборник статей

по материалам VI Всероссийской научно-практической конференции, Москва, 12 апреля 2023 года. – Москва : Российский государственный гуманитарный университет, 2023. – С. 97–101.

22. Рябченко, Н. А. Применение методов сетевого анализа и компьютерной лингвистики для анализа деструктивного контента в сети Интернет / Н. А. Рябченко, О. П. Малышева, С. В. Усков // MEDIAОбразование: цифровая среда в условиях вынужденной метаморфозы : сборник материалов VII Международной научно-практической конференции, Челябинск, 22–24 ноября 2022 года. – Челябинск : Челябинский государственный университет, 2022. – С. 367–372.

23. Первый философ Искусственного Интеллекта // Хабр. – URL: https://habr.com/ru/articles/694934/ (дата обращения: 29.09.2023).

24. McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955 // AI Magazine. – 2006. – № 27 (4). – P. 12. https://doi.org/10.1609/aimag.v27i4.1904.

25. ChatGPT: China detains man for allegedly generating fake train crash news, first known time person held over use of AI bot // SCMP. – URL: https://www.scmp.com/news/china/politics/article/3219764/china-announces-first-known-chatgpt-arrest-over-alleged-fake-train-crash-news (дата обращения: 01.10.2023).

26. Garbage in, garbage out // Wikipedia. – URL: https://en.wikipedia.org/wiki/Garbage_in,_garbage_out (дата обращения: 01.10.2023).

References

1. DIGITAL 2023: GLOBAL OVERVIEW REPORT. *datareportal*. URL: https://datareportal.com/reports /digital-2023-global-overview-report (accessed 08.08.2023).

2. Kara-Murza, S. G. *Manipulation of consciousness*, 2003. 464 p. URL: http://lib.ru/POLITOLOG /karamurza.txt (accessed 10.08.2023).

3. Krasovskaya, N. R., Gulyaev, A. A. To the issue of classification of information wars. *Sociology of Science and Technology*, 2019, vol. 10, no. 2, pp. 44–55. DOI: 10.24411/2079-0910-2019-12002.

4. *Article 10.1. Obligations of the organiser of information dissemination in the Internet*. URL: https://www.consultant.ru/document/cons_doc_LAW_61798/9ab7abe2b9fe407f507610f7e6e14a951d575585/ (accessed 13.08.2023).

5. *Article 15.1. Unified register of domain names, page indexes of Internet sites and network addresses allowing to identify Internet sites containing information the dissemination of which is prohibited in the Russian Federation*. URL: https://www.consultant.ru/document/cons_doc_LAW_61798/38c8ea666d27d9dc12b078c556e316e90248f551 /#dst16 (accessed 13.08.2023).

6. Big Encyclopaedic Dictionary. *Wikipedia*. URL: https://ru.wikipedia.org/wiki/Большой _энциклопедический_словарь (accessed 21.08.2023).

7. Ivin, A. A., Nikiforov, A. L. *Dictionary of Logic*. Moscow, VLADOS Humanitarian Publishing Centre, 1997. 384 pp.

8. Philosophical Encyclopaedic Dictionary. *Dictionaries Online*. URL: https://rus-philosophy-enc. slovaronline.com/ (accessed 21.08.2023).

9. Collection of encyclopaedias and dictionaries of the Russian language. *Meaning of the word Credibility in the encyclopaedic dictionary*. URL: https://diclist.ru/slovar/enciklopedicheskiy/d/1-dostovernost.html (accessed 21.08.2023).

10. Collection of encyclopaedias and dictionaries of the Russian language. *Meaning of the word Credibility in the dictionary of logic.* URL: https://diclist.ru/slovar/logiki/d/dostovernost.html (accessed 21.08.2023).

11. Collection of encyclopaedias and dictionaries of the Russian language. *Meaning of the word Credibility in the philosophical dictionary*. URL: https://diclist.ru/slovar/logiki/d/dostovernost.html (accessed 21.08.2023).

12. Big explanatory sociological dictionary. *gufo.me*. URL: https://gufo.me/dict/social_dict?page=5&letter=д (accessed 21.08.2023).

13. Big explanatory sociological dictionary. *Dictionaries online.* URL: https://rus-philosophy-enc. slovaronline.com/ (accessed 21.08.2023).

14. NBC suspends journalist Brian Williams for 'lying' about brushing with death in Iraq war. *SCMP*. URL: https://www.scmp.com/news/world/article/1709945/nbc-suspends-journalist-brian-williams-lying-about-brush-death-iraq-war (accessed 19.09.2023).

15. NBC TV host apologised for lying on air and interrupted work. *RIA Novosti*. URL: https://ria.ru/20150208/1046536663.html?ysclid=ln0b9d9wgm318468016 (accessed 19.09.2023).

16. How TV channel «Imedi» «killed» the President of Georgia. *Lenta.ru.* URL: https://lenta.ru/articles/2010/03/15/imedi/] (accessed 20.09.2023).

17. Israel bombs as tanks wait on outskirts of Gaza. *CNN.* URL: https://edition.cnn.com/2009/WORLD/meast/01/02/israel.gaza/index.html (accessed 20.09.2023).

18. French aviation museum played a trick on journalists by promising to fly Concorde. *RIA Novosti*. URL: https://ria.ru/20090401/166698857.html?ysclid=lmz0yz41n5231955908] (accessed 21.09.2023).

19. Karabak, I. I., Zorin, K. A., Azhmukhamedov, I. M. *Certificate of State Registration of Computer Programme no. 2021680437 Russian Federation. Programme for parsing messages in public telegram channels* ; no. 2021669987 ; appl. 30.11.2021 ; publ. on 10.12.2021.

20. Minaev, V. A., Simonov, A. V., Rebrova, A. D. Automated detection of destructive content in social media. *Information security: yesterday, today, tomorrow : collection of articles on the materials of the IV International Scientific and Practical Conference, Moscow, 22 April 2021*. Moscow, Russian State University for the Humanities, 2021, pp. 124–130.

21. Veprev, S. B., Nesterovich, S. A. Application of Big Data technology to identify destructive content on the Internet. *Information security: yesterday, today, tomorrow : collection of articles on the materials of the VI All-Russian scientific-practical conference, Moscow, 12 April 2023.* Moscow, Russian State University for the Humanities, 2023, pp. 97–101.

22. Ryabchenko, N. A., Malysheva, O. P., Uskov, S. V. Application of methods of network analysis and computational linguistics for the analysis of destructive content on the Internet. *MEDIAeducation: digital environment in conditions of forced metamorphosis : proceedings of the VII International Scientific and Practical Conference, Chelyabinsk, 22–24 November 2022.* Chelyabinsk, Chelyabinsk State University, 2022, pp. 367–372.

23. The first philosopher of Artificial Intelligence. *Hubr.* URL: https://habr.com/ru/articles/694934/ (accessed 29.09.2023).

24. McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine*, 2006, no. 27 (4), p. 12. https://doi.org/10.1609/aimag.v27i4.1904.

25. ChatGPT: China detains man for allegedly generating fake train crash news, first known time person held over use of AI bot. *SCMP.* URL: https://www.scmp.com/news/china/politics/article/3219764/china-announces-first-known-chatgpt-arrest-over-alleged-fake-train-crash-news (accessed 01.10.2023).

26. Garbage in, garbage out. *Wikipedia.* URL: https://en.wikipedia.org/wiki/Garbage_in,_garbage_out (accessed 01.10.2023).

Статья поступила в редакцию 27.09.2023; одобрена после рецензирования 12.10.2023; принята к публикации 12.10.2023.

The article was submitted 27.09.2023; approved after reviewing 12.10.2023; accepted for publication 12.10.2023.

DOI 10.54398/20741707_2023_4_36
УДК 004.056

АЛГОРИТМ КОНТРОЛЯ ЭКСФИЛЬТРАЦИИ ДАННЫХ С УЧЕТОМ ТРЕБОВАНИЙ УПРАВЛЕНИЯ НА ОСНОВЕ ДАННЫХ

Золотарев Вячеслав Владимирович, Сибирский государственный университет науки и технологий, 660037, Российская Федерация, г. Красноярск, пр. им. газ. «Красноярский рабочий», 31,

кандидат технических наук, доцент, заведующий кафедрой безопасности информационных технологий, ORCID: 0000-0002-8054-8564, e-mail: zolotorev@sibsau.ru

При реализации работы с данными в задаче управления на основе данных возникают новые проблемы управления информационной безопасностью. Они должны быть решены через создание применимых в указанной задаче алгоритмов, моделей, методик и подходов управления безопасностью, в том числе на уровне организации процессов, работы с данными и формирования архитектуры информационной безопасности организации. Изучение вопросов управления информационной безопасностью для задачи управления на основе данных должно также лежать в основе развития современных систем обмена и агрегирования данных. Настоящая работа содержит описание проблемы эксфильтрации данных, описывает уровни и возможные варианты ее решения. Представлены общие алгоритм и подходы к решению задачи поиска нетегированных конфиденциальных данных при оценке возможности эксфильтрации данных. Схема, предложенная в работе, может быть использована как для имитационных моделей, так и для реализации в виде набора процессов управления информационной безопасностью в практических задачах.

Ключевые слова: управление информационной безопасностью, процессный подход, алгоритм управления безопасностью, образовательный процесс, управление на основе данных, эксфильтрация данных

Финансирование: *исследование выполнено при финансовой поддержке Минцифры РФ (грант ИБ, проект № 40469-01/2022-д).*

DATA EXFILTRATION CONTROL ALGORITHM WITH THE REQUIREMENTS OF DATA-BASED MANAGEMENT

Zolotarev Vyacheslav V., Siberian State University of Science and Technology, 31 Krasnoyarsky Rabochy Ave., Krasnoyarsk, 660037, Russian Federation,

Cand. Sci. (Engineering), Associate Professor, Head of Information Technologies Security Department, ORCID: 0000-0002-8054-8564, e-mail: zolotorev@sibsau.ru

When implementing work with data in a data-based management task, new problems of information security management arise. They should be solved through the creation of algorithms, models, methods and approaches of security management applicable to this task, including at the level of organizing processes, working with data and forming the organization's information security architecture. The study of information security management issues for the task of data-based management should also underlie the development of modern data exchange and aggregation systems. This paper describes the problem of data exfiltration, describes the levels and possible solutions to it. The general algorithm and approaches to solving the problem of searching for untagged confidential data when assessing the possibility of data exfiltration are presented. The scheme proposed in the paper can be used both for simulation models and for implementation as a set of information security management processes in practical tasks.

Keywords: information security management, process approach, security management algorithm, educational process, information infrastructure, data-based management, data exfiltration

Financial support: *the work was supported by the Ministry of Digital Development, Communications and Mass Communications of the Russian Federation (IS grant, project No. 40469-01/2022-D).*

Graphical annotation (Графическая аннотация)

ВВЕДЕНИЕ

Управление информационной безопасностью должно учитывать работу с данными различной степени конфиденциальности, в том числе и с непомеченными данными, потенциально требующими ограничения доступа. Известны решения по поиску и тегированию конфиденциальной информации, к примеру, в патентах [1–2] и ряде смежных работ предполагается следующий результат подобного поиска – повышение точности классификации конфиденциальной информации за счет обработки полученных данных с помощью ансамбля нейронных сетей, в ходе которой данным в каждой ячейке таблицы присваивается тег, соответствующий заданному типу конфиденциальной информации. При этом предобработка текстовой информации предполагает приведение ее к виду, пригодному для анализа, в частности векторного представления [3].

Также проблема работы с данными при управлении на основе данных, поиска конфиденциальной информации в базах данных рассматривалась в работах российских ученых [4–6].

Проблема же наличия и распространения непомеченных данных в рамках настоящей работы следует из особенностей управления на основе данных, рассмотренных далее.

Проблема **эксфильтрации данных** возникает при обмене данными между участниками процессов и объектами, содержащими необходимые им данные (узлами документооборота, внешними и внутренними агрегированными базами данных). При этом возникающие потоки данных могут быть нарушающими принятые политики безопасности в отношении информационного обмена, а коммуникации – скрытыми, неявными или слабоконтролируемыми [7].

При реализации управления на основе данных добавляется сложность контроля автоматизированных и автоматических коммуникаций и коннекторов, существующих как элемент автоматизации бизнес-процессов и роботизированных программных решений [8–10]. Эти коммуникации могут быть либо не полностью контролируемыми с позиций распространения конфиденциальной информации (или иных типов информации, контроль над которыми может представлять интерес для организации), либо находиться в состоянии неполного контроля в определенных переходных состояниях организационных и информационных систем [11].

Таким образом, для управления на основе данных должны существовать отдельные правила и применимые политики, исключающие или снижающие вероятность эксфильтрации данных.

Исследование, предлагаемое вниманию, содержит рекомендации по внедрению алгоритмов контроля, по введению строгих и слабых (основанных на риск-ориентированном подходе) требований к контролю эксфильтрации данных, что придает гибкость задачам управления безопасностью и работы с информационными ресурсами организации.

На уровне концепции существует проблема следующего вида: на уровне управления на основе данных существует необходимость получить доступ к как можно большему количеству источников для сбора, анализа, валидации и верификации информации; с точки зрения информационной безопасности это будет нарушением принципа минимизации полномочий и локализации объектов защиты информации в инфраструктуре организации, если не будут приняты соответствующие усилия по формированию дополнений в политику информационной безопасности организации. Причина такого противоречия в том, что эффективность управления с увеличением количества источников данных должна возрастать [12], а сложность реализации систем защиты информации для подобных решений может становиться неприемлемой из-за возможных неполных или нецелесообразных подходов и решений.

Проблема эксфильтрации данных на различных уровнях управления на основе данных (см. также [4]) может выглядеть следующим образом (рис.). На схеме ниже показаны основные проблемы эксфильтрации данных для этапов работы с ними в рамках управления информационной безопасностью.

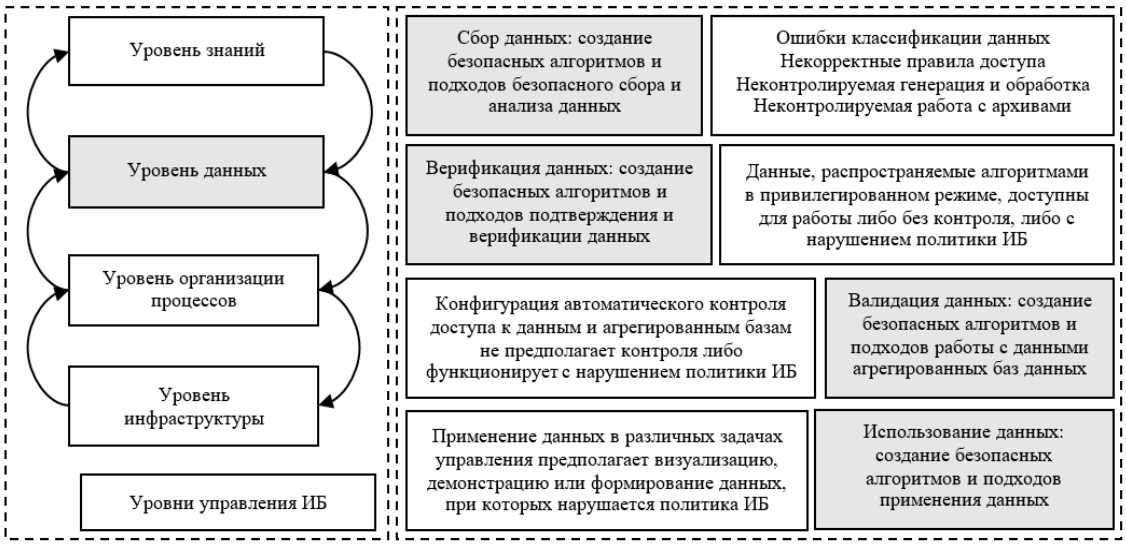


Рисунок – Проблемы эксфильтрации данных для управления на основе данных

Итак, далее рассмотрена на уровне алгоритма работы с тегированными данными и поиска данных, потенциально требующих ограничения доступа, задача контроля эксфильтрации данных для отдельных задач, указанных на схеме выше.

Представляется важным также учесть изменение требований к системе защиты информации для новых процессов для будущего практического применения.

Целью исследования, результаты которого приведены ниже, является решение следующей задачи: предложить алгоритм и модель управления эксфильтрацией данных с учетом требований к управлению на основе данных.

Новыми результатами исследования, представленными ниже, стали:

1) конкретизация ранее предложенной авторами четырехуровневой модели управления информационной безопасностью, включающей слой управления знаниями, для описания процессов управления информационной безопасностью, затрагивающих несколько уровней этой модели (а именно процессов, работающих для уровня данных);

2) формирование новых устойчивых связей поддерживающих процессов управления информационной безопасностью на уровне отдельных задач с учетом четырехуровневой модели с акцентом на уровень данных.

ПРОЦЕСС КОНТРОЛЯ ЭКСФИЛЬТРАЦИИ ДАННЫХ

Управление информационной безопасностью в целом должно генерировать базовые процессы, применимые в задаче [13], такие как указанные ниже (табл. 1). Процессы должны учитывать как требования локальных политик безопасности, так и наборы ограничений, формируемые в рамках управления требованиями в системе управления информационной безопасностью организации.

Исходное условие контроля эксфильтрации, выведенное в статье [14], предполагает следующее:

Представим обработку данных при принятии управленческого решения в виде системы упорядоченных векторов: $S = (s_0, s_1, \dots, s_n)$, где n – количество выполненных процедур, s_0 – начальное состояние, s_n – заключительное состояние.

Для поиска уязвимостей процессов определим D как множество всех агрегированных баз данных, доступных для обращения, в том числе ретроспективно. Также определим $D_i = f(s_i)$ как множество результатов запросов к фрагментам агрегированных баз данных, доступных для обращения из состояния s_i и $D_i \subseteq D, i \in [0; n]$. f задана алгоритмически.

В свою очередь, пусть имеется Dt_i – множество доступных агрегированных баз данных для обращения из состояния s_i и $Dt_i \subseteq D$. Пусть $Dp_i \subseteq D$ некоторое подмножество фрагментов этих баз, хранящих защищаемую информацию в рамках состояния s_i.

Обозначим через γ функцию, которая позволяет определить, является ли состояние s_i допустимым или недопустимым, следующим образом:

$$\gamma(s_i, D) = \begin{cases} 2, \text{если } D_i \cap (Dp_i \backslash Dt_i) \neq \emptyset \\ 1, \text{если } ((D_i \cap Dp_i = \emptyset) \wedge (\mathrm{D_i} \cap (\mathrm{D} \backslash \mathrm{Dt_i}) \neq \emptyset)) \vee (D_i \nsubseteq D) \\ 0, \text{во всех других случаях} \end{cases} \quad (1)$$

2 – недопустимое состояние и есть уязвимость процесса;

1 – недопустимое состояние и есть ошибка регламента;

0 – допустимое состояние.

Таблица 1 – Применимые (и недостающие) процессы управления ИБ для уровня данных

Процесс	Назначение	Требования	Примечание
Сбор данных	Создание безопасных алгоритмов и подходов безопасного сбора и анализа данных	Ограничение доступа к данным (не все данные должны подлежать автоматическому или автоматизированному сбору, ограничение по ролям и обязанностям) Срок хранения и место хранения, а также условия хранения должны быть определены	В том числе для внешних источников
Верификация данных	Создание безопасных алгоритмов и подходов подтверждения и верификации данных	Ограничение доступа к источникам вспомогательных данных для верификации (записям, документам, регистрационным данным) Ограничение доступа к данным для случая запроса внешних заинтересованных лиц Ограничение доступа к данным для случая запроса внутренних заинтересованных лиц	В том числе в рамках повторной верификации и для работы с неполными (поврежденными, остаточными) данными
Валидация данных	Создание безопасных алгоритмов и подходов работы с данными агрегированных баз данных	Ограничение доступа к тестовым и имитационным моделям Ограничение доступа к источникам вспомогательных данных для валидации (записям, документам, регистрационным данным)	В том числе для сложнофункциональных тестовых и имитационных моделей, таких как цифровые двойники объектов
Использование данных	Создание безопасных алгоритмов и подходов применения данных	Управление доступом к данным на всех этапах жизненного цикла	В том числе при уничтожении данных
Общие процессы		Ограничение доступа к данным, используемым для формирования моделей чувствительных бизнес-процессов Учет требований при формировании руководящих указаний в части информационной безопасности	

Там же представлен и риск-ориентированный подход к оценке, при котором распространение данных предполагается с определенной вероятностью; следовательно, контроль эксфильтрации должен быть реализован для узлов, вероятность неконтролируемого распространения данных для которых выше. Оценка такой вероятности, по мнению авторов, является отдельной задачей, заслуживающей внимания.

При этом минимальный уровень риска задается в качественном виде, т. е. рассматривается категория ценности фрагмента данных и вероятность его эксфильтрации, а не количественные значения ценности (что является отдельной сложной задачей).

АЛГОРИТМ КОНТРОЛЯ ЭКСФИЛЬТРАЦИИ

Алгоритм контроля эксфильтрации в такой задаче может быть реализован следующим образом:

0. Формирование признаков данных, позволяющих выполнить предобработку (на уровне работы с метаданными, местом расположения данных, временем их создания или изменения, электронных подписей данных или их хеш-значений).

1. Данные, выделенные как контролируемые, должны быть размечены тем или иным образом. Предполагаемая разметка должна быть машиночитаемой и применяться универсальным образом.

2. Размеченные данные анализируются на наличие существенных признаков.

2 (а). Выполняется сортировка данных, в том числе и для вновь получаемых данных в автоматических и автоматизированных алгоритмах сбора и обработки данных, с учетом выделенных признаков и разметки данных.

В качестве алгоритмов разметки данных могут быть использованы следующие варианты (табл. 2).

Таблица 2 – Действия с помеченными данными

Признак	Порядок работы	Требования	Примечание
Шаблон конфиденциальной информации	Сравнение с заранее сгенерированным или полученным из тестовой выборки шаблоном	Шаблон должен быть определен Шаблон должен покрывать все виды анализируемых данных для всех источников данных Шаблон должен быть пригоден для сопоставления или формирования дополнительных признаков для эффективного анализа	В том числе для внешних источников
Результат классификации или кластеризации	Обучение на размеченных данных (тестовом примере) и формирование набора признаков на основе обучения	Наличие размеченной выборки Обновление размеченной выборки для повторного обучения (в случае инцидентов с эксфильтрацией данных) Учет вероятностного характера оценки	Требует вычислительных ресурсов
Контент	Анализ содержимого контейнера, содержащего данные (файла, пакета, дампа трафика и т. п.)	Наличие применимых алгоритмов для всех видов контролируемого контента	
Хеш-значение	Сравнение заранее сгенерированных хеш-значений для контейнеров, содержащих контролируемые данные (файлов, пакетов, дампов трафика и т. п.) с анализируемыми контейнерами, содержащими данные	Наличие образцов для сравнения Безопасное хранение образцов для хранения Архитектура системы, учитывающая требования безопасного хранения образцов	Требует развертывания системы хранения данных для безопасного хранения образцов
Метаданные	Сравнение и анализ метаданных файлов (названия, авторы, даты создания и т. п.) для оценки риска наличия конфиденциальной или иной контролируемой информации в нем	Учет вероятностного характера оценки Наличие метаданных Наличие применимых алгоритмов для всех видов контента	

3. В зависимости от процесса должно быть выполнено следующее корректирующее действие (табл. 3). Суть корректирующего действия должна быть в выявлении признаков данных, контроль эксфильтрации которых представляет интерес для организации, а также выявления нарушений применимых политик и анализе потоков данных в организации.

Таблица 3 – Корректирующие действия

Процесс	Действие алгоритма	Результат	Примечание
Сбор данных	Анализ помеченных данных	Выявление существенных признаков данных, контроль эксфильтрации которых представляет интерес для организации	В том числе для внешних баз данных и внешних связей организации
	Работа с тестовыми данными	Выявление нарушений применимых политик для потоков данных организации	В том числе для внешних баз данных и внешних связей организации
	Построение схемы потоков данных	Гиперграф на основе организационных схем и диаграмм потоков данных	В том числе с учетом существующих формализованных и неформализованных способов обработки данных
	Определение доступных мест хранения и обработки данных	Уточненная диаграмма потоков данных	
Верификация данных	Анализ состояний s_i	Выявление процессов с уязвимостями и ошибками в политиках обработки данных	В том числе в рамках повторной верификации и для работы с неполными (поврежденными, остаточными) данными
	Определение ранее не определенных (редких, исключительных, вновь возникающих) состояний s_i	Выявление процессов с уязвимостями и ошибками в политиках обработки данных Уточнение множества всех агрегированных баз данных, доступных для обращения, в том числе ретроспективно доступных	
	Анализ вновь выявленных состояний s_i	Выявление процессов с уязвимостями и ошибками в политиках обработки данных	В том числе в рамках повторной верификации и для работы с неполными (поврежденными, остаточными) данными
Валидация данных	Формирование схем сравнения данных тестовых моделей (имитационных моделей, цифровых двойников)	Протоколы доступа к тестовым моделям (имитационным моделям, цифровым двойникам) для контролируемых данных Корректировка применимых локальных и глобальных политик безопасности	
	Определение ранее не определенных (редких, исключительных, вновь возникающих) состояний s_i	Выявление процессов с уязвимостями и ошибками в политиках обработки данных Уточнение множества всех агрегированных баз данных, доступных для обращения, в том числе ретроспективно доступных, с учетом множества агрегированных баз данных тестовых моделей (имитационных моделей, цифровых двойников)	
	Анализ вновь выявленных состояний s_i	Выявление процессов с уязвимостями и ошибками в политиках обработки данных	В том числе в рамках повторной валидации (например, при ошибке доступа) и для работы с неполными (поврежденными, остаточными) данными

3 (а). Уточняющие (корректирующие) действия. В зависимости от результатов контроля могут быть также выполнены определенные корректировки применимых политик и алгоритмов обработки, хранения и анализа данных. Выше (табл. 3) также показаны этапы, на которых могут быть применены подобные корректирующие действия.

На уровне использования данных применимые действия должны зависеть от вида и способа обработки контролируемых данных и определяются отдельно.

4. Формирование итоговых оценок эксфильтрации данных для выявленных состояний.

5. Анализ итоговых оценок и формирование рекомендаций.

Таким образом, в рамках формирования алгоритмической основы должны быть выявлены и проанализированы определенные состояния системы, потенциально уязвимые с точки зрения проблемы эксфильтрации данных, а также проанализированы возможности контроля эксфильтрации с позиций применимых практик и политик в области информационной безопасности.

Проблема эксфильтрации данных может усугубляться при использовании данных, не имеющих тегирования на этапе создания или формируемых в ходе работы системы; для таких данных должен быть предусмотрен алгоритм контроля эксфильтрации, предполагающий выявление их в ходе работы системы.

Примеры такого алгоритма для данных, имеющих известные признаки (шаблоны), по которым возможен контроль эксфильтрации, представленных в текстовом виде, аудиоинформации или иных пригодных для контентного анализа видах, приведены для DLP-систем (систем предотвращения утечек конфиденциальной информации).

Далее же рассматривается ситуация, когда таких признаков заранее не могло быть сформировано.

ЭКСПЕРИМЕНТ ПО КОНТРОЛЮ ЭКСФИЛЬТРАЦИИ ДАННЫХ

Принцип эксперимента в следующем: проверяется возможность оценки эксфильтрации данных для ситуации, в которой заранее размеченных данных и шаблонов для разметки получено не было. Был сформирован тестовый набор данных, содержащий конфиденциальную информацию в виде фото и сканированных изображений, контрольный набор данных для обучения с размеченными данными. Инструментом работы с данными была выбрана нейронная сеть, использующая обучение без учителя, и предложен следующий алгоритм работы с данными:

1. Подготовка данных: собрать набор данных, состоящий из файлов изображений различных типов, которые содержат или не содержат конфиденциальную информацию.

1 (а). Для обучения в рамках контрольного набора данных файлы изображений размечены соответствующим образом для выделения конфиденциальной информации (экспертным путем).

2. Построение модели: формируется архитектура нейронной сети, которая будет обрабатывать входные файлы и выдавать предсказания о наличии или отсутствии конфиденциальной информации в каждом файле. При этом предпочтительным способом анализа является анализ на основе метаданных.

3. Обучение модели: на подготовленном наборе данных выполняется обучение нейронной сети. Оценивается ошибка прогноза о наличии конфиденциальной информации в файле.

4. Тестирование модели: выполняется тестирование на отдельном наборе данных, которые не использовались в процессе обучения. Оценивается точность и производительность полученного решения.

Если данные успешно **выделены и помечены**, далее должен использоваться следующий **алгоритм работы с помеченными данными**, контролирующий их эксфильтрацию вне зависимости от уровня применимой политики или процесса:

1. Для помеченных (контролируемых) данных анализ предполагает сопоставление с шаблоном или набором признаков, полученных нейросетью, и формирование соответствующей метки (тега, метаданных изображения).

2. Помеченные (контролируемые) данные должны храниться, обрабатываться и передаваться только теми компонентами системы, которые отмечены как обрабатывающие контролируемые данные. Если такой информации о данном компоненте нет, то:

2 (а). Вносится соответствующая отметка

или

2 (б). Прекращается обработка (хранение, передача) контролируемых данных.

3. Формируется (корректируется) инструкция по работе с контролируемыми данными.

Следовательно, для каждого случая выявления эксфильтрации данных должен быть зафиксирован инцидент информационной безопасности, и обработка этого инцидента должна завершиться либо изменением списка компонентов системы, где возможна обработка контролируемых данных, либо выявлением и контролем эксфильтрации контролируемых данных.

ЗАКЛЮЧЕНИЕ

Исследование сосредоточено на формировании пригодных в практике рекомендаций по построению процессной модели управления информационной безопасности, учитывающей как традиционно рассматриваемые уровни управления инфраструктурой информационной системы и управления организационными процессами, так и новые, расширяющие модель уровни данных, в том числе вопрос неконтролируемого распространения (эксфильтрации) данных, рассмотренный в настоящей работе.

Подобное расширение может быть полезно при развертывании разных типов экспертных и советующих систем, систем поддержки принятия решений, ситуационных центров и особенно интересно для задач управления знаниями, опирающихся на описанный уровень данных, которые возможно рассмотреть в отдельных исследованиях.

Кроме того, интерес представляет задача поиска нетегированных (неразмеченных) конфиденциальных данных в открытых базах данных в том случае, если данные не поддаются эффективному контентному анализу.

Список источников

1. Способ и система классификации данных для выявления конфиденциальной информации : патент RU 2 759 786 C1 / Алексей Алексеевич Теренин. – Заявл. 05.07.2019 ; опубл. 17.11. 2021.

2. Способ и система классификации данных для выявления конфиденциальной информации в тексте : патент RU 755 606 C2 / Алексей Алексеевич Теренин. – Заявл. 16.10.2019 ; опубл. 17.09.2021.

3. Способ и система получения векторного представления электронного текстового документа для классификации по категориям конфиденциальной информации : патент RU 2 775 358 C1 / Кирилл Евгеньевич Вышегородцев. – Заявл. 24.09.2021 ; опубл. 29.06.2022.

4. Золотарев, В. В. Модель и алгоритм управления информационной безопасностью образовательной организации высшего образования с учетом требований управления на основе данных / В. В. Золотарев, М. А. Лапина // Прикаспийский журнал: управление и высокие технологии. – 2022. – № 4 (60). – С. 107–118. – DOI: 10.54398/20741707_2022_4_107.

5. Martishin, S. A. Study of the problem of ensuring security in the storage and processing of confidential data / S. A. Martishin, M. V. Khrapchenko, A. V. Shokurov // Труды ИСП РАН. – 2021. – Т. 33, вып. 2. – С. 173–190. – DOI: 10.15514/ISPRAS–2021–33(2)–11.

6. Хоанг, В. К. Метод контроля прямого доступа к семантическим базам данных / В. К. Хоанг, А. Ф. Тузовский // Известия Томского политехнического университета. – 2013. – Т. 322, № 5. – С. 138–142.

7. ГОСТ Р 53113.2-2009. Национальный стандарт Российской Федерации. Информационная технология. Защита информационных технологий и автоматизированных систем от угроз информационной безопасности, реализуемых с использованием скрытых каналов. – Москва : Росстандарт, 2009.

8. Алаудинов, А. Г. Интеграция данных корпоративных информационных систем / А. Г. Алаудинов // Качество. Инновации. Образование. – 2011. – № 11 (78). – С. 64–69.

9. Белякова, О. А. Платформа для сбора, хранения и анализа мультимодальных данных, генерируемых субъектами онлайн коммуникаций / О. А. Белякова, П. Н. Махнин, С. В. Сапегин // Мониторинг. Наука и технологии. – 2019. – № 4 (42). – С. 73–84. – DOI: 10.25714/MNT.2019.42.011.

10. Васильев, Ю. С. Обеспечение безопасности автоматизированных систем управления технологическими процессами на объектах гидроэнергетики / Ю. С. Васильев, П. Д. Зегжда, Д. П. Зегжда // Известия Российской академии наук. Энергетика. – 2016. – № 3. – С. 49–61.

11. Золотарев, В. В. Моделирование и расчет параметров управления информационной безопасностью переходных состояний цифровой трансформации образовательного учреждения высшего образования / В. В. Золотарев, П. А. Звягинцева // Решетневские чтения : материалы XXVI Международной научно-практической конференции, посвященной памяти генерального конструктора ракетно-космических систем академика М. Ф. Решетнева : в 2 ч., Красноярск, 09–11 ноября 2022 года / под общ. ред. Ю. Ю. Логинова. – Красноярск : Федеральное государственное бюджетное образовательное учреждение высшего образования «Сибирский государственный университет науки и технологий имени академика М. Ф. Решетнева», 2022. – Ч. 2. – С. 396–398.

12. Бабикова А. В. Информационный ресурс как инструмент повышения эффективности управления корпоративной компанией // Известия ЮФУ. Технические науки. – 2006. – № 10. – URL: https://cyberleninka.ru /article/n/informatsionnyy-resurs-kak-instrument-povysheniya-effektivnosti-upravleniya-korporativnoy-kompaniey (дата обращения: 23.09.2023).

13. Hashim, R. Contributing Factors for Successful Information Security Management Implementation: A Preliminary Review / R. Hashim, R. Rozilawati // Internal journal of innovative technology and exploring engineering. – 2019. – Vol. 9, issue 2, December

14. Zolotarev, V. Data Exfiltration Control in the Data-Driven Management / V. Zolotarev, M. Lapina, E. Lapina, M. Anita, M. Sajid // Adv. in Systems Science and Appl. (in print).

References

1. Terenin, Alexey Alekseevich. *Method and system of data classification for identifying confidential information : patent RU 2,759,786 C1.* Appl. 05.07.2019 ; publ. 17.11. 2021.

2. Terenin, Alexey Alekseevich. *Method and system of data classification for identifying confidential information in the text : patent RU 755 606 C2.* Appl. 16.10.2019 ; publ. 17.09.2021.

3. Vyshegorodtsev, Kirill Evgenievich. *Method and system for obtaining a vector representation of an electronic text document for classification by categories of confidential information : patent RU 2,775,358 C1*. Appl. 24.09.2021 ; publ. 29.06.2022.

4. Zolotarev, V. V., Lapina, M. A. Model and algorithm of information security management of an educational organization of higher education taking into account the requirements of data-based management. *Caspian Journal: Control and High Technologies*, 2022, no. 4 (60), pp. 107–118. DOI: 10.54398/20741707_2022_4_107.

5. Martishin, S. A., Khrapchenko, M. V., Shokurov, A. V. Investigation of the problem of security in the storage and processing of confidential data. *Proceedings of ISP RAS*, 2021, vol. 33, issue 2, pp. 173–190. DOI: 10.15514/ISPRAS–2021–33(2)–11.

6. Hoang, V. K., Tuzovsky, A. F. Method of controlling direct access to semantic databases. *Proceedings of Tomsk Polytechnic University*, 2013, vol. 322, no. 5, pp. 138–142.

7. *GOST R 53113.2-2009. National Standard of the Russian Federation. Information technology. Protection of information technologies and automated systems from information security threats implemented using hidden channels.* Moscow, Rosstandart Publ., 2009.

8. Alaudinov, A. G. Data integration of corporate information systems. *Quality. Innovation. Education*, 2011, no. 11 (78), pp. 64–69.

9. Belyakova, O. A., Makhnin, P. N., Sapegin, S. V. Platform for collection, storage and analysis of multimodal data generated by subjects of online communications. *Monitoring. Science and technology*, 2019, no. 4 (42), pp. 73–84. DOI: 10.25714/MNT.2019.42.011.

10. Vasiliev, Yu. S., Zegzhda, P. D., Zegzhda, D. P. Ensuring the safety of automated process control systems at hydropower facilities. *News of the Russian Academy of Sciences. Energy*, 2016, no. 3, pp. 49–61.

11. Zolotarev, V. V., Zvyagintseva, P. A. Modeling and calculation of information security management parameters of transitional states of digital transformation of an educational institution of higher education. *Reshetnev Readings : materials of the XXVI Scientific and Practical International Conference dedicated to the memory of the general designer of rocket and space systems Academician M.F. Reshetnev*, 2 parts, Krasnoyarsk, November 09–11, 2022. – Krasnoyarsk, Federal State Budgetary Educational Institution of Higher Education "Siberian State University of Science and Technology named after Academician M.F. Reshetnev", 2022, part 2, pp. 396–398.

12. Babikova, A.V. Information resource as a tool to improve the efficiency of corporate company management. *News of the SFU. Technical sciences*, 2006, no. 10. URL: https://cyberleninka.ru/article/n/informatsionnyy-resurs-kak-instrument-povysheniya-effektivnosti-upravleniya-korporativnoy-kompaniey (accessed 09.23.2023).

13. Hashim, R. Factors contributing to the successful implementation of information security management: a preliminary review / R. Hashim, R. Rozilavati. *Internal Journal of Innovative Technologies and Research Engineering*, 2019, vol. 9, issue 2, December.

14. Zolotarev, V. Lapina, M., Lapina, E., Anita, M., Sajid, M. Control of data exfiltration in data-based management. *Adv. in Systems Science and Appl.* (in print).

Статья поступила в редакцию 02.10.2023; одобрена после рецензирования 09.10.2023; принята к публикации 16.10.2023.

The article was submitted 02.10.2023; approved after reviewing 09.10.2023; accepted for publication 16.10.2023.

МАТЕМАТИЧЕСКОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ, КОМПЛЕКСОВ И КОМПЬЮТЕРНЫХ СЕТЕЙ

УДК 004.891.2

РАЗРАБОТКА МОДУЛЯ ДЛЯ ОПТИМИЗАЦИИ РАСЧЕТА ОПТИЧЕСКОЙ СИЛЫ ИНТРАОКУЛЯРНОЙ ЛИНЗЫ

Тарапатина Екатерина Сергеевна, Волгоградский государственный технический университет, 400005, Российская Федерация, г. Волгоград, пр. им. В.И. Ленина, 28,
магистрант, ORCID: 0009-0001-9591-4731, e-mail: taras320@mail.ru
Виноградов Артем Рудольфович, Волгоградский филиал ФГАУ НМИЦ МНТК «Микрохирургия глаза» им. акад. С.Н. Федорова, 400138, Российская Федерация, г. Волгоград, ул. им. Землячки, 80,
врач-офтальмолог, ORCID: 0009-0001-9591-4731, e-mail: art_vino@mail.ru
Сибирный Никита Денисович, Волгоградский государственный технический университет, 400005, Российская Федерация, г. Волгоград, пр. им. В.И. Ленина, 28,
аспирант, ORCID: 0009-0000-4653-4495, e-mail: lockk1337@gmail.com
Орлова Юлия Александровна, Волгоградский государственный технический университет, 400005, Российская Федерация, г. Волгоград, пр. им. В.И. Ленина, 28,
доктор технических наук, доцент, ORCID: 0000-0003-4854-7462, e-mail: yulia.orlova@ gmail.com
Фролов Максим Юрьевич, Волгоградский государственный медицинский университет, 400131, Российская Федерация, г. Волгоград, площадь Павших Борцов, 1,
кандидат медицинских наук, доцент, ORCID: 0000-0003-2679-7524, e-mail: clinpharmrussia @yandex.ru

Данное исследование посвящено описанию модуля для оптимизации расчета оптической силы интраокулярной линзы. Этот модуль будет предоставлять расчет оптической силы ИОЛ на основе данных, предоставленных Волгоградским филиалом ФГАУ НМИЦ МНТК «Микрохирургия глаза» имени академика С.Н. Федорова, для хирургического лечения катаракты, а также хранения показаний исследования пациента. Система будет включать в себя модуль медицинской информационной системы МНТК, REST-сервер, SPA-приложение, а также модуль искусственного интеллекта, который позволит определить наилучший показатель силы ИОЛ для данного пациента.
Ключевые слова: медицинская информационная система, SPA-приложение, интраокулярная линза, катаракта, REST-сервер

DEVELOPMENT OF A MODULE FOR OPTIMIZING THE CALCULATION OF THE OPTICAL POWER OF AN INTRAOCULAR LENS

Tarapatina Ekaterina S., Volgograd State Technical University, 28 V.I. Lenin Ave., Volgograd, 400005, Russian Federation,
master's student, ORCID: 0009-0001-9591-4731, e-mail: taras320@mail.ru
Vinogradov Artem R., Volgograd branch of the Federal State Autonomous Institution National Medical Research Center MNTK "Eye Microsurgery" named after academician S.N. Fedorov, 80 Zemlyachki St., Volgograd, 400138, Russian Federation,
ophthalmologist, ORCID: 0009-0001-9591-4731, e-mail: art_vino@mail.ru
Sibirny Nikita D., Volgograd State Technical University, 28 V.I. Lenin Ave., Volgograd, 400005, Russian Federation,
graduate student, ORCID: 0009-0000-4653-4495, e-mail: lockk1337@gmail.com
Orlova Yulia A., Volgograd State Technical University, 28 V.I. Lenin Ave., Volgograd, 400005, Russian Federation,
Doct. Sci. (Engineering), Assistant Professor, ORCID: 0000-0003-4854-7462, e-mail: yulia. orlova@gmail.com
Frolov Maxim Yu., Volgograd State Medical University, Volgograd, 1 Fallen Fighters Square, 400131, Russian Federation,
Cand. Sci. (Medicine), Assistant Professor, ORCID: 0000-0003-2679-7524, e-mail: clinpharmrussia @yandex.ru

This study is devoted to the description of a module for optimizing the calculation of the optical power of an intraocular lens. This module will provide the calculation of the optical strength of the intraocular lens based on data provided by the Volgograd branch of the FSAU NMIC MNTC "Eye Microsurgery" named after academician S.N. Fedorov, for the surgical treatment of cataracts, as well as the storage of patient examination indications. The system will include a MNTC medical information system module, a REST server, a SPA application, as well as an artificial intelligence module that will determine the best indicator of IOL strength for a given patient.
Keywords: medical information system, SPA application, intraocular lens, cataract, REST server

ВВЕДЕНИЕ

При проведении офтальмологических операций врачам необходимо провести большое количество различных исследований как до операции, так и после. Это необходимо для того, чтобы правильно определить стратегию проведения операции и получить подтверждение правильности выбора после операции.

Катаракта – заболевание глаз, которое характеризуется помутнением хрусталика [1]. Из этого вытекает одна из самых основных проблем офтальмологии – значительное увеличение заболеваемости ею [2]. Выявление данного заболевания обычно происходит на его поздней стадии, когда терапевтическое лечение уже малоэффективно. Поэтому необходимо как можно быстрее назначить пациенту правильное лечение и прооперировать его, вернув ему таким образом нормальное зрение.

Перед непосредственной операцией хирургу необходимо выбрать правильную силу интраокулярной линзы (ИОЛ) для того, чтобы получить максимальный рефракционный результат у пациента. Механизм выбора ИОЛ не формализован до конца, обычно этот выбор производится на основе клинического опыта самого офтальмохирурга и его коллег. Но если попытаться выделить алгоритм действий, то он будет выглядеть следующим образом:

– расчет значений ИОЛ по всем доступным формулам и нахождение среднее из них;

– выявление особенностей у пациента. Например, при коротких глазах (менее 22 мм длиной) наиболее подходящей формулой является Hoffer Q. У данной формулы есть свой эмпирический коэффициент ошибки – 0,5 дптр. А если взять пациентов с длинными глазами, то им лучше уже подойдут формулы Barrett или Kane;

– выявление особых персональных потребностей пациента.

Офтальмохирург (непосредственно человек, который занимается вычислением ИОЛ) тратит достаточно много времени на его подбор для каждого пациента. Исходя из этого, можно заключить, что данному алгоритму необходима автоматизация путем внедрения модуля с машинными обучением, который позволил бы рассчитывать ИОЛ всеми актуальными на данный момент средствами и, основываясь на предыдущих решениях в подобных случаях, рекомендовал врачу наиболее подходящие значения.

Цель данного исследования – это создание системы поддержки принятия врачебного решения (СППВР) для расчета оптической силы ИОЛ для хирургического лечения катаракты, что позволит повысить точность, а также уменьшить время расчета ИОЛ, разгрузив таким образом время врача.

АНАЛИЗ ПРЕДМЕТНОЙ ОБЛАСТИ

За последнее десятилетие различные высокотехнологичные ИОЛ вошли в ежедневную практику в катарактальной хирургии, доказав возможность эффективного восстановления зрительных функций [7]. Вместе с этим существует большая проблема структуризации и хранения медицинской информации. Так, например, авторы исследования [2], проведя обзор состояния и развития этой области, выяснили, что с внедрением клинических информационных систем увеличивается процент полноты и корректность клинических и лабораторных данных. Исходя из этого, можно составить концептуальную диаграмму предметной области (рис. 1).

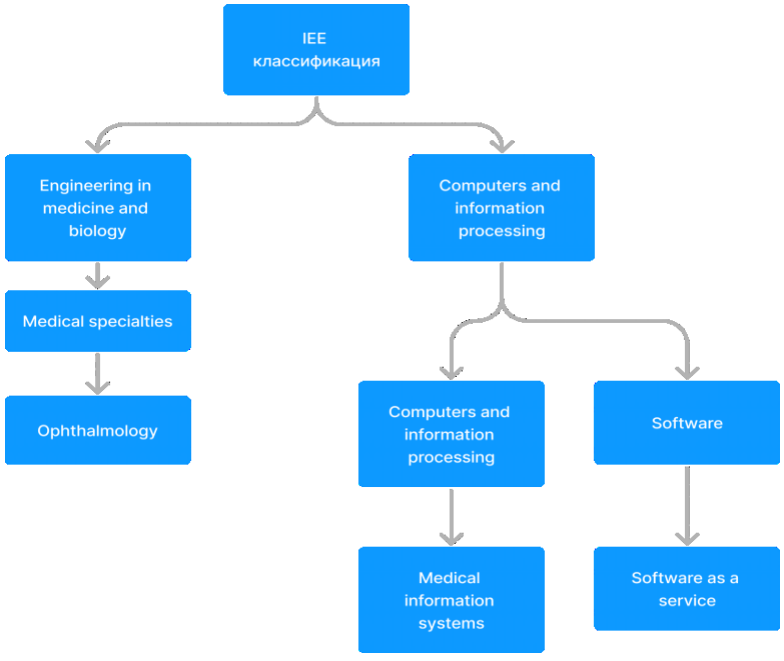


Рисунок 1 – ER-диаграмма объектной модели предметной области

Рассмотрим текущий процесс вычисления ИОЛ, представленный во введении (рис. 2).

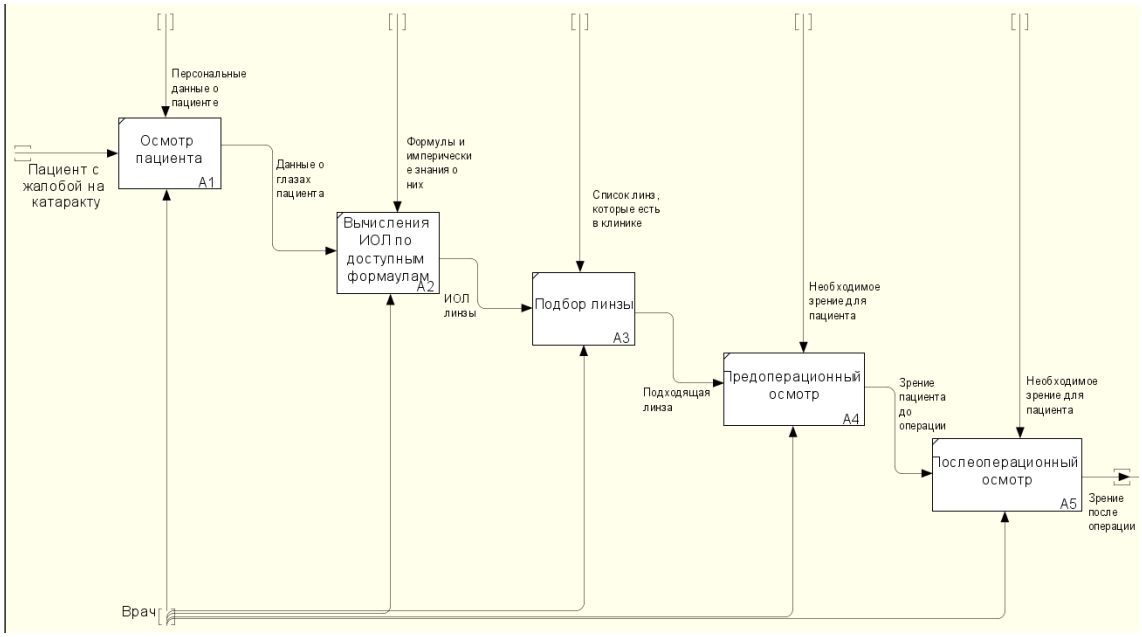


Рисунок 2 – Диаграмма бизнес-процесса выбора линзы для пациента до внедрения модуля

Здесь мы можем видеть, что наиболее трудоемкие этапы, такие как вычисление ИОЛ по доступным формулам или подбор ИОЛ, выполняются офтальмохирургом. Оба этапа (вычисление ИОЛ по доступным формулам и подбор линзы) предлагается оптимизировать путем внедрения модуля вычисления ИОЛ. Это позволит сэкономить время врача для принятия правильного решения. Процесс после оптимизации представлен на рисунке 3.

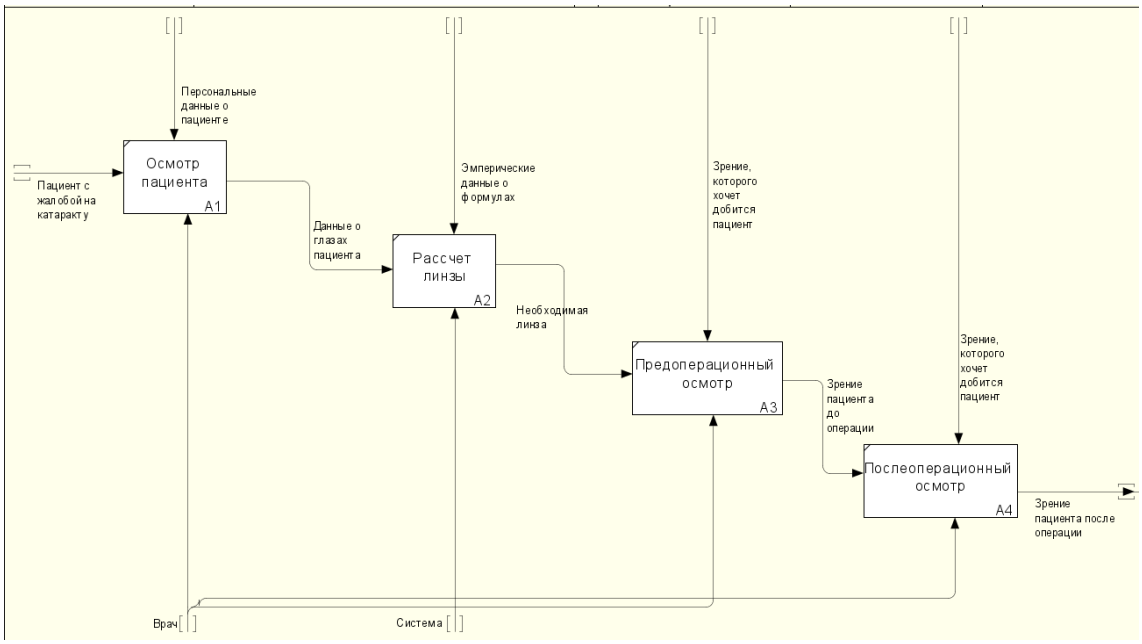


Рисунок 3 – Диаграмма бизнес-процесса выбора линзы для пациента после внедрения модуля

ПЛАН ИССЛЕДОВАНИЯ

Как было сказано ранее, на данный момент наиболее остро стоит вопрос о цифровизации медицины. В данном исследовании [8] говорится о проблеме систематизации и формализации медицинской информации, где предложили модель кодификации клинической информации. В заключении данного исследования авторы утверждают, что кодификация информации позволит решить проблему кореферентности. Поэтому необходимо эффективно собирать и анализировать данные для вычислений.

Основываясь на этом, вначале нашего исследования необходимо собрать информацию обо всех исследованиях, которые проводятся перед операцией. После необходимо проанализировать все ключевые методы расчета ИОЛ, которые наиболее популярны в практике врачей РФ, далее необходимо понять, какие замеры проводятся до и после операции для оценки ее эффективности.

Следующим этапом будет разработка программной системы, которая будет в автоматическом режиме собирать перспективные данные о пациентах до и после операции: кератометрия, рефрактометрия и др. Помимо этого, собираться будут также данные о ИОЛ, выбранной для операции (А-константа, Barrett LF, диоптрийность ИОЛ и т.д.), а также выбранная офтальмохирургом формула для расчета ИОЛ, которая использовалась при проведении данной операции. В ходе данной разработки будет разработан UI/UX-дизайн SPA-приложения, само SPA-приложение (web-клиент данного модуля) и REST-сервер, а также архитектура БД.

После этого начнется внедрение данного модуля в медицинскую информационную систему (МИС) федерального государственного автономного учреждения «Национальный медицинский исследовательский центр» межотраслевого научно-технического комплекса «Микрохирургия глаза» имени академика С.Н. Федорова (МНТК). В ходе этого этапа SPA-приложение, REST-сервер и БД будут размещены на серверах МНТК. Помимо этого, будет разработан и внедрен алгоритм синхронизации данных МИС МНТК с данным модулем.

Четвертым этапом исследования будет набор данных для анализа и обучения математической модели. Сами данные будут набираться в автоматическом режиме благодаря интеграции с МИС МНТК на третьем этапе. Помимо набора данных, будет производиться их анализ и удаление выбросов для корректного обучения модели.

Пятый этап – подбор и обучение ML-модели. Предполагается, что модель будет представлять из себя классификатор по выбору необходимой линзы для пациента. После этого она будет интегрирована в существующую цепочку SPA-приложения, REST-сервера и МИС МНТК.

Итогом данного исследования должна стать СППВР, которая будет основываться на предыдущих операционных вмешательствах для лечения катаракты и проведенных расчетах для подбора ИОЛ.

Для оценки достижения цели данная СППВР пройдет тестирование на данных МНТК, а также впоследствии будет апробирована и протестирована медицинскими экспертами для выявления точности. Тестирование будет проходить следующим образом: будет сравниваться целевое зрение пациента и полученное зрение в результате операции с вычисленной ИОЛ. Помимо зрения, будут оцениваться возможные осложнения, вызванные вычисленной ИОЛ, а также удовлетворенность пациента текущей ИОЛ (на одном из послеоперационных осмотров пациенту будет предложено пройти тестирование для оценки сервиса).

Диаграмма последовательности этапов исследования представлена на рисунке 4.

Рисунок 4 – Диаграмма последовательности этапов исследования

ОПИСАНИЕ РАЗРАБАТЫВАЕМОГО ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

Разработанный модуль предназначен для использования в соответствии с рядом функциональных требований:

– просмотр и корректировка результатов расчета ИОЛ;

– расчет и принятия решения по использованию соответствующей ИОЛ;

– синхронизация данных с существующей МИС.

Необходимыми требованиями к разрабатываемой системе являются требования законодательства, нормативных актов по защите персональных и медицинских данных (медицинских баз данных), а также к хранению медицинской информации. Исходя из этого, диаграмма вариантов использования представлена на рисунке 5.

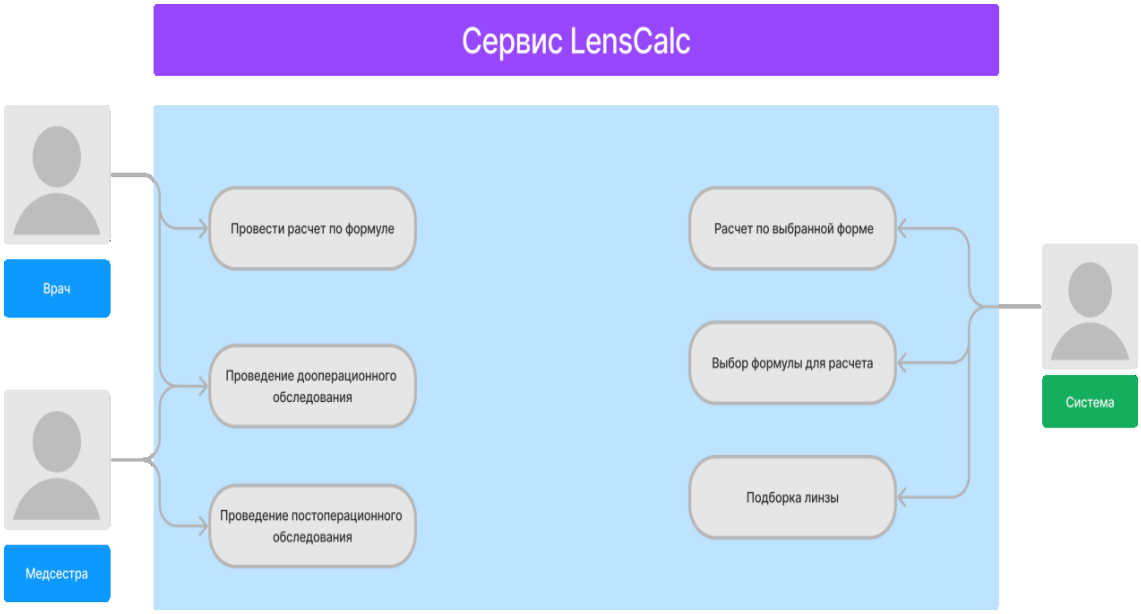


Рисунок 5 – Диаграмма вариантов использования

Диаграмма взаимодействия основных компонентов данного модуля представлена на рисунке 6.

Одним из основных элементов данного модуля является модель, которая будет написана на Python и обернута в микросервис в виде REST-сервера на Flask. Помимо этого, будет и основной сервис, являющийся REST-сервером, логика которого написана на Java под управлением фреймворка Spring Boot. Все данные сервисов будут хранится в БД под управлением СУБД MySQL.

Клиентское приложение будет представлять собой SPA-приложение, которое написано на JavaScript с использованием фреймворка Vue.

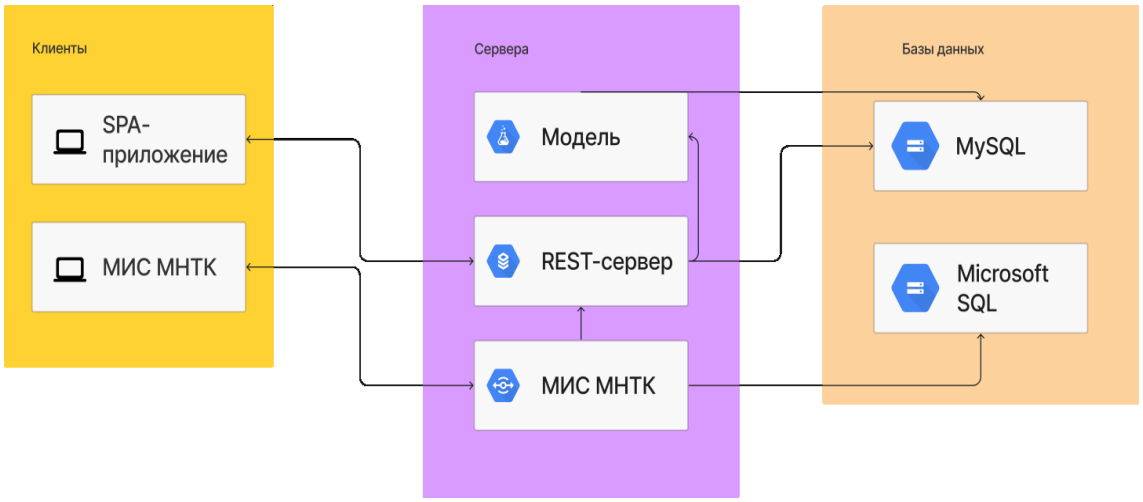


Рисунок 6 – Описание информационной системы

Помимо этого, REST-сервер синхронизирован с данными из МИС МНТК посредством REST-запросов.

На рисунке 7 изображена схема БД. В данной схеме appointment – это таблица с исследованиями пациента при посещении врача, где врач лично осматривает пациента. Данные с приборов хранятся в четырех таблицах – device, parameter, parameter_value и unit_parameter_value. Device и parameter – таблицы, в которых хранятся данные о том, какие параметры принадлежат конкретному прибору, а parameter_value и unit_parameter_value – это таблицы, в которых лежат уже непосредственные значения этого прибора. Таблицы formula, formula_parameter и unit_calculation – хранение значений формул (formula), их параметров (formula_parameter) и рассчитанных по ним значений (unit_calculation). Линзы и их константы, в свою очередь, хранятся в таблицах lenses и constants соответственно.

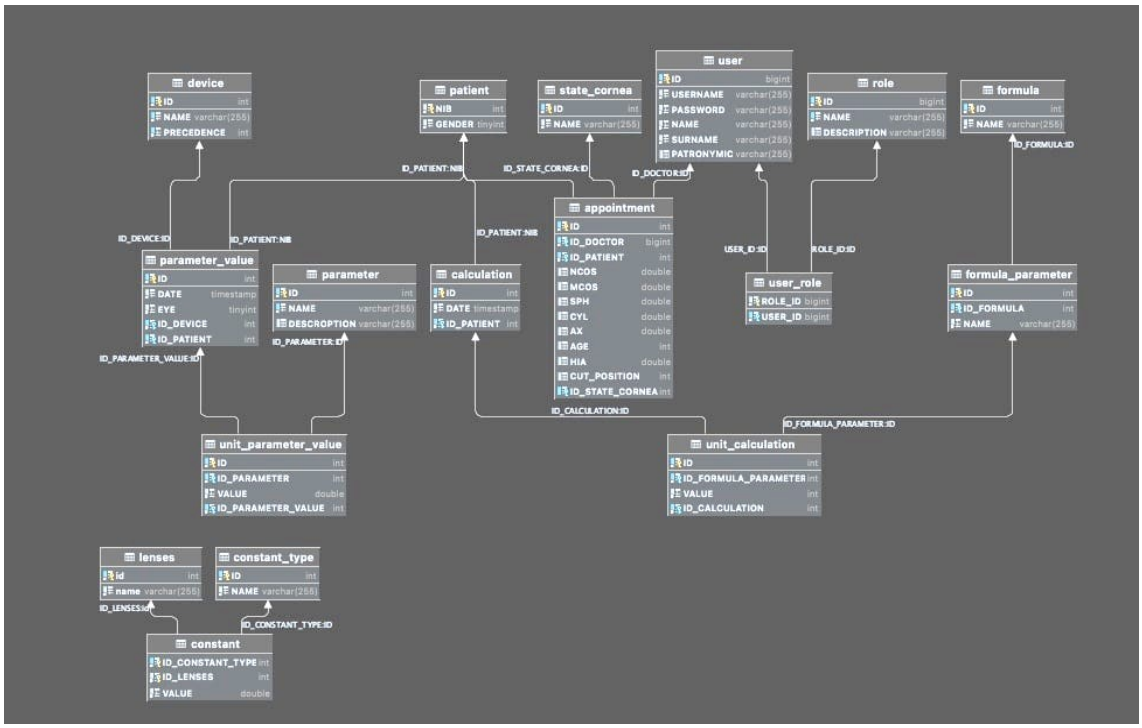


Рисунок 7 – Схема БД

ОПИСАНИЕ МОДУЛЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Одним из наиболее значимых элементов данного модуля является искусственный интеллект, включающий в себя расчет по заданным формулам, а также выбор наиболее подходящего значения для данного случая катаракты.

Для обучения модели, которая в дальнейшем позволит проводить вычисления ИОЛ, необходимо выделить следующие фичи, содержащиеся в формулах Barrett Suite, Haigis, Holladay 2, Hoffer Q, Kane, Barrett TrueK Toric, которые представлены в таблице.

Предполагается, что тип модели в данном сервисе будет классификатором, который будет «классифицировать» для определенного случая катаракты свою ИОЛ. Входные данные для этой модели являются признаковым описанием, которое является в том числе и фичи. Типы классов будут варьироваться, предполагается, что это будет многоклассовая классификация, но не исключаются и пересекающиеся классы (случаи, когда человеку могут подойти несколько линз).

Таблица – Выделенные признаки для обучения модели машинного обучения для определения класса ИОЛ

Название характеристики	**Пример данных (в чем измеряется)**
SE по выбранной формуле	–0,21
Выбранная диоптрийность ИОЛ	+21,5
Угол имплантации ИОЛ	95
Остаточная сфера	–0,11
Остаточный цилиндр	–0,20
Угол остаточного цилиндра	185
А-константа ИОЛ	119,13
Длина глаза (AL)	24,5
Глубина передней камеры (ACD)	3,5
Толщина хрусталика (LT)	4,5
K1 и ее AX	44,1
K2 и ее AX	90
CCT	535
PK1 и ее AX	–5,6
PK2 и ее AX	–5,6

По мере расширения и обучения модели данные фичи будут ранжироваться или упразднятся.

Предполагается, что целевой результат – это причисление к одному или нескольким классам из имеющихся в клинике ИОЛ.

РАСЧЕТ ПО ЗАДАННЫМ ФОРМУЛАМ

Для того чтобы выяснить, какая именно нам нужна ИОЛ, ее необходимо рассчитать по определенным формулам. В данном модуле будет реализован расчет по следующим формулам: Barrett (Universal II, Toric, TrueK), Haigis, Holladay 2, Hoffer Q, Kane и Barrett TrueK Toric. Данные формулы имеют разную погрешность при вычислении ИОЛ, поэтому на практике лучше использовать вычисления нескольких формул.

Вообще все формулы нового поколения не показали систематической ошибки, в то время как все традиционные формулы имеют статистически значимый сдвиг от 0,36 до 0,65 дптр. в глазах с высокой миопией, за исключением Haggis (0,28 дптр).

ДАЛЬНЕЙШЕЕ РАСШИРЕНИЕ СИСТЕМЫ

Рассматриваемый модуль в дальнейшем будет расширен и доработан до системы поддержки принятия решений для расчета силы ИОЛ для дальнейшего проведения офтальмологических операций по хирургическому лечению катаракты. В систему будут добавлены: новые методы классификации; будет дальнейшая доработка системы в сторону универсальности интеграции установки системы в различные медицинские центры, которые занимаются проведением операций по лечению катаракты.

ЗАКЛЮЧЕНИЕ

В продолжении грядущего исследования будет разработан модуль, обеспечивающий повышение точности и уменьшение времени принятия решений при расчете ИОЛ для хирургического лечения катаракты, дальнейшая ее апробация в МНТК на предмет ошибок и корректной работы модуля.

Список источников

1. Покровский, В. И. Малая медицинская энциклопедия / В. И. Покровский. – Москва : Советская энциклопедия, 1991. – Т. 2. – 624 с.

2. Затрудина, Р. Ш. Современные методы диагностики катаракты на ранней стадии развития / Р. Ш. Затрудина, К. С. Ивина, Н. В. Марусин // NBI-technologies. – 2012. – № 6.

3. Кузьминов, О. М. Проблема кореферентности и модель кодификации клинической информации / О. М. Кузьминов, Н. Н. Шаламова, О. В. Муромцева // Актуальные проблемы медицины. – 2013. – № 11 (154).

4. Богуш, И. В. Комбинированный метод определения оптической силы интраокулярных линз после кератотомии / И. В. Богуш // Сибирский научный медицинский журнал. – 2009. – № 4.

5. Зенкова, Н. А. Возможности компьютерного и математического моделирования для решения некоторых задач офтальмологии / Н. А. Зенкова // Вестник российских университетов. Математика. – 2017. – № 6–2.

6. Демченко, М. В. Разработка медицинской информационной системы с элементами поддержки принятия решений в кардиологии / М. В. Демченко, М. А. Фирюлина, И. Л. Каширина // МНИЖ. – 2021. – № 8–1 (110).

7. Морозова, Т. А. Современные аспекты мультифокальной интраокулярной коррекции (обзор) / Т. А. Морозова, Д. Ф. Покровский, И. Б. Медведев, Т. З. Керимов // Вестник РАМН. – 2017. – № 4.

8. Старичкова, Ю. В. Подходы к интеграции комплекса программных средств управления процессами и сложноструктурированными медицинскими данными с медицинскими и лабораторными информационными системами в учреждениях здравоохранения / Ю. В. Старичкова, К. А. Воронин, Н. В. Борисова, М. А. Масчан, А. Г. Румянцев // Известия вузов. Поволжский регион. Технические науки. – 2016. – № 4 (40).

References

1. Pokrovsky, V. I. *Small medical encyclopedia*. Moscow, Soviet Encyclopedia, 1991, vol. 2. 624 p.

2. Peremodina, R. S., Ivina, K. S., Marusin, N. V. Modern methods of cataract diagnosis at an early stage of development. *NBI-technologies*, 2012, no. 6.

3. Kuzminov, O. M., Shalamova, N. N., Muromtseva, O. V. The problem of coreference and the model of codification of clinical information. *Actual problems of medicine*, 2013, no. 11 (154).

4. Bogush, I. V. Combined method for determining the optical strength of intraocular lenses after keratotomy. *Siberian Scientific Medical Journal*, 2009, no. 4.

5. Zenkova, N. A. Possibilities of computer and mathematical modeling for solving some problems of ophthalmology. *Bulletin of Russian Universities. Mathematics*, 2017, no. 6–2.

6. Demchenko, M. V., Firyulina, M. A., Kashirina, I. L. Development of a medical information system with elements of decision support in cardiology. *MNIZH*, 2021, no. 8–1 (110).

7. Morozova, T. A., Pokrovsky, D. F., Medvedev, I. B., Kerimov, T. Z. Modern aspects of multifocal intraocular correction (review). *Bulletin of the Russian Academy of Medical Sciences*, 2017, no. 4.

8. Starichkova, Yu. V., Voronin, K. A., Borisova, N. V., Maschan, M. A., Rumyantsev, A. G. Approaches to the integration of a complex of software tools for managing processes and complex structured medical data with medical and laboratory information systems in healthcare institutions. *News of universities. Volga region. Technical sciences*, 2016, no. 4 (40).

Статья поступила в редакцию 15.05.2023; одобрена после рецензирования 05.06.2023; принята к публикации 23.06.2023.

The article was submitted 15.05.2023; approved after reviewing 05.06.2023; accepted for publication 23.06.2023.

DOI 10.54398/20741707_2023_4_52
УДК 004.001

ДИАГНОСТИКА ТУБЕРКУЛЕЗА БЕЗ БАКТЕРИОВЫДЕЛЕНИЯ С ПРИМЕНЕНИЕМ КЛАССИЧЕСКИХ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Тюлькова Татьяна Евгеньевна, Федеральное государственное бюджетное учреждение «Национальный медицинский исследовательский центр фтизиопульмонологии и инфекционных заболеваний» Министерства здравоохранения Российской Федерации, г. Москва (ФГБУ «НМИЦ ФПИ» Минздрава России), 127473, Российская Федерация, Москва, ул. Достоевского, 4,

доктор медицинских наук, руководитель отдела координации научных исследований, ORCID: 0000-0002-2292-1228), e-mail: tulkova2006@rambler.ru

Чернавин Павел Федорович, Уральский федеральный университет, 620002, Российская Федерация, г. Екатеринбург, ул. Мира, 19,

кандидат экономических наук, доцент, ORCID: 0000-0003-3214-3906, e-mail: chernavin.p.f@gmail.com

Чернавин Николай Павлович, Уральский федеральный университет, 620002, Российская Федерация, г. Екатеринбург, ул. Мира, 19,

ассистент, ORCID: 0000-0002-2093-9715, e-mail: ch_k@mail.ru

Постановка диагноза по туберкулезу в случае отсутствия бактериовыделения является сложной и неоднозначной задачей даже для фтизиатров с большим стажем. В данной статье исследуются 139 входных параметров, учитываемых при постановке диагноза о наличии туберкулеза. Оперировать таким количеством признаков при постановке диагноза сложно даже профессионалам. Поэтому в ходе человеко-машинной процедуры делается оценка информативности признаков и их количество снижается до 62. После этого профессиональные фтизиатры разбивают их на шесть групп исходя из собственного опыта и нормативных требований. По каждой группе решаются задачи классификации семью классическими методами машинного обучения. Результаты разбиения на группы и метрики качества различных методов приведены в таблицах. Статья представляет интерес как для профессиональных фтизиатров, так и специалистов по машинному обучению.

Ключевые слова: туберкулез без бактериовыделения, машинное обучение, информативность признаков

TUBERCULOSIS DIAGNOSIS WITHOUT BACTERIAL EXCRETION USING CLASSICAL METHODS OF MACHINE LEARNING

Tyulkova Tatyana E., Federal State Budgetary Institution “National Medical Research Center for Phthisiopulmonology and Infectious Diseases” of the Ministry of Health of the Russian Federation, Moscow (FSBI "NMIC FPI" of the Ministry of Health of Russia), 4 Dostoevsky St., Moscow, 127473, Russian Federation,

Doct. Sci. (Medicine), Head of the Department of Scientific Research Coordination, ORCID: 0000-0002-2292-1228), e-mail: tulkova2006@rambler.ru

Chernavin Pavel F., Ural Federal University, 19 Mira St., Yekaterinburg, 620002, Russian Federation,

Cand. Sci. (Economics), Associate Professor, ORCID: 0000-0003-3214-3906, e-mail: chernavin.p.f@ gmail.com

Chernavin Nikolai P., Ural Federal University, 19 Mira St., Yekaterinburg, 620002, Russian Federation,

assistant, ORCID: 0000-0002-2093-9715, e-mail: ch_k@mail.ru

Diagnosing tuberculosis in the absence of bacterial excretion is a complex and ambiguous task even for phthisiatricians with extensive experience. This article examines 139 input parameters taken into account when diagnosing tuberculosis. Managing such a number of characteristics while diagnosing is challenging even for professionals. Therefore, during the human-machine procedure, an assessment of the informativeness of the features is made and their number is reduced to 62. After this, professional phthisiatricians divide them into six groups based on their own experience and regulatory requirements. Classification tasks are solved for each group using seven classical machine learning methods. The results of the division into groups and the quality metrics of various methods are presented in tables. The article is of interest both for professional phthisiatricians and specialists in machine learning.

Keywords: tuberculosis without bacterial excretion, machine learning, feature informativeness

ВВЕДЕНИЕ

Обнаружение возбудителя Mycobacterium tuberculosis (МБТ) или его ДНК в биологическом материале пациентов оценивается как верифицированный диагноз туберкулеза, а установленным он считается при наличии клинико-рентгенологических признаков заболевания, отсутствии бактериовыделения и/или гистологического подтверждения диагноза [8]. Основной причиной отсутствия возбудителя туберкулеза в респираторном материале является незначительное его количество в очаге воспаления при небольшом участке поражения или наличие препятствия (анатомического или функционального) между очагом воспаления и внешней средой, а также отсутствия диагностического материала (мокроты) для проведения исследования.

Для такого типа клинических ситуаций диагностика строится на основе методов, позволяющих косвенно судить об инфекции. В некоторых ситуациях требуется биопсия (операция) с целью получения материала для последующих исследований. В случае отсутствия возможности и/или отсутствия результата на подтверждение/опровержение туберкулеза в силу технических причин, возникает необходимость проводить терапию ex yuvantibus. В мире ежегодно проводится свыше 26,5 млн курсов терапии «отчаяния» [10]. При этом врачу приходится анализировать множество клинических признаков, в том числе уточнять жалобы, подозрительные на туберкулез, а также лабораторные показатели и результаты иммунологических и рентгенологических методов обследования [1, 7, 12]. Столь значительное количество вводной информации затрудняет их интерпретацию, приводит к отсроченной диагностике (нередко основывающейся на эффектах терапии ex yuvantibus). Это приводит к увеличению страданий пациента, нерациональному использованию лекарственных препаратов, прием которых может привести к изменению параметров пациента и усугубить его состояние. Одним из путей решения этой проблемы является построение моделей диагностики с использованием искусственного интеллекта, который позволяет анализировать большое количество переменных.

Таким образом, **целью** нашего исследования стал поиск комбинации показателей для математической обработки и построения модели диагностики туберкулеза на основе комплексной оценки клинико-рентгенологических и лабораторных показателей с применением методов искусственного интеллекта.

МАТЕРИАЛЫ И МЕТОДЫ

В анализ включены 139 показателей, на основании которых была заполнена база данных (свидетельство о регистрации базы данных № 2021621556 от 09.07.2021 г.).

Среди них были жалобы, которые были структурированы по принадлежности к синдрому интоксикации (снижение массы тела, утомляемость при выполнении повседневной нагрузки, повышение температуры и т. д.) и респираторному синдрому (кашель и его длительность, одышка и т. д.); демографические и антропометрические параметры с указанием пола, возраста, веса, роста с расчетом индекса массы тела; региона проживания; социального статуса и группы риска, длительности заболевания до выявления, а также метод выявления; факт наличия вредных привычек в настоящее время или давностью до 12 месяцев; факт установленного контакта и его варианты; информация о пребывании в местах лишения свободы и давность проведения флюорографического обследования; факт проведения микробиологического обследования в медицинских организациях первичного звена здравоохранения и фибробронхоскопия (ФБС); результаты обследования на туберкулез (микробиологические и молекулярно-генетические); диагноз при выявлении туберкулеза с указанием формы, фазы, распространенности процесса, осложнений, в том числе неспецифических, поражение бронхов, а также сопутствующей патологии и группы учета; наличие и вид оперативного вмешательства; описание сопутствующей патологии с уточняющими данными по сахарному диабету.

Факт наличия сопутствующей патологии регистрировался при указании на него в данный момент или давностью не более 1 года. В базу данных были внесены данные о лучевых методах обследования в виде рентгенсимптомов, объединенных в укрупненные группы, о результатах гистоморфологического обследования (при наличии) и лабораторных исследований при первичном обследовании и в динамике через 2–4 недели.

СТАТИСТИЧЕСКИЕ МЕТОДЫ ОБРАБОТКИ ДАННЫХ

Статистическая обработка данных проводилась с помощью прикладных программ «Microsoft Excel 97», БИОСТАТИСТИКА для Windows, SPSS – 14.0, предназначенных для выполнения всех этапов статистического анализа, согласуемого с принципами клинической эпидемиологии «evidence based medicine». Использованы два вида данных: дискретные (типа да/нет) и интервальные (количественные показатели). Описание количественных показателей выполнено с помощью среднего арифметического значения (M) ± стандартное отклонение (δ), характеризовавшего выборку внутри группы. Рассчитывалась доля описываемого признака. Различия интервальных переменных в независимых выборках анализировали с помощью двухвыборочного *t*-критерия Стьюдента с поправкой на неравенство дисперсий (наличие разнородности внутри группы) по Levene. Статистически значимыми считали различия при $p < 0,05$. Для оценки дискретных признаков и встречаемости изучаемого фактора применяли отношение шансов (OR). Для регрессионного анализа применяли OR и ДИ. Величину различий оценивали путем расчета разности средних и определения 95 % доверительного интервала (ДИ). Если ДИ этой разности не содержал внутри себя 1, то гипотеза о равенстве средних отвергалась. Использован расчет площади под ROC-кривой

При неправильном распределении использовался непараметрический критерий Манна – Уитни (U) для проверки гипотезы о равенстве средних в группах. Гипотеза отвергалась при значении $p < 0,05$.

Для определения принадлежности пациентов с ограниченным процессом в легких к больным туберкулезом применялся регрессионный метод. Дискриминантный анализ помогал осуществлять интегрированную оценку клинико-эпидемиологических, лабораторных методов обследования и с достаточной точностью классифицировать пациентов. Для всех 139 показателей базы данных были рассчитаны индексы информативности признаков на основе функции f-classif из библиотеки Sklearn.

Согласуясь с мнением экспертов, данных литературы, нормативно-правовой документации

[1–6, 8, 10–12], были сформированы 6 комбинаций показателей, претендующих на диагностическую математическую модель. То есть с точки зрения научного подхода (в философском смысле) были сформированы 6 гипотез для построения решающих правил и требовалось различными методами машинного обучения оценить каждую из них.

При этом была использована следующая процедура. Если мнение экспертов о низкой информативности признака совпадало с оценкой функцией f-classif, то данный признак исключался из рассмотрения. В противном случае признак оставляли как претендента на участие в модели. Таким образом количество признаков удалось сократить до 62.

Информативные признаки и их комбинации приведены в таблице 1.

Таблица 1 – Комбинации признаков для включения в диагностическую модель

Показатели			**f-classif**	**Варианты комбинаций**					
				I	**II**	**III**	**IV**	**V**	**VI**
Возраст (лет)			31,59	V	V	V	V	V	V
Пол			0,90	V		V	V	V	
Социальный статус			320,52	V		V			V
Группы риска		Фтизиатрическая	97,20	V	V	V	V		V
		Отсутствие флюорографии ≥2 лет	0,40	V	V	V			
		Социальная	20,45	V					V
		Медицинская	0,68	V					V
Вредные привычки		Алкоголь	145,52	V					V
		Наркотики	0,01	V					
		Курение	0,01	V					
Способ выявления		При обращении	8,43	V	V	V			V
		При профосмотре	42,11	V	V	V			V
Факт нахождения в местах лишения свободы			0,58	V					
Жалобы на момент обращения	Интоксикационный синдром	Снижение массы тела	8,69	V		V			V
		Потливость	1,43	V					
		Температура более 37	0,36	V		V			
		Слабость при выполнении прежней нагрузки	8,45	V					V
	Респираторный синдром	Кашель	12,4	V		V			V
		Кровохарканье	25,11	V					
		Боли в груди	4,99	V					
Сопутствующие заболевания по факту выявления без уточнения нозологии			45,13	V	V				
Вирусный гепатит С			1,51						V
Вирусный гепатит В			3,29						V
Гнойные инфекции кожи, подкожной клетчатки, органов брюшной полости, суставов			1,95						V
Патология ЛОР-органов			10,17						V

Показатели		f-classif	Варианты комбинаций					
			I	II	III	IV	V	VI
Патология желудочно-кишечного тракта		2,46						V
Патология почек или мочевыделительной системы		0,18						
Перенесенная пневмония		32,60						V
Урогенитальные заболевания		1,19						V
Дисбактериоз		0,17						
Острые респираторные инфекции		1,95						V
Атопический дерматит		0,52						
Бронхиальная астма		5,93						V
Онкологические заболевания		129,72						V
Психические заболевания		0,58						
Ревматоидный артрит		6,42						V
Патология щитовидной железы		4,38						V
Хроническая обструктивная болезнь легких		4,84						V
Патология сердечно-сосудистой системы		30,12						V
Сахарный диабет I / II типа		2,40						V
Рентгенологические симптомы								
Очаговое образование / несколько очагов		0,003					V	
Деструкция		0,28					V	
Утолщение плевры/спайки		20,45					V	
Кальцинаты		0,12					V	
Усиленный сосудистый рисунок		6,72					V	
Лабораторные показатели								
Результаты анализа крови	Rbc (эритроциты)	0,60	V		V	V	V	
	WBC (лейкоциты)	24,45	V	V	V	V	V	V
	PLT (тромбоциты)	3,58				V	V	V
	HGB (гемоглобин)	5,30	V	V	V	V	V	V
	Эозинофилы	14,47				V	V	V
	Палочкоядерные нейтрофилы	19,84	V	V	V	V	V	V
	Сегментоядерные нейтрофилы	10,47	V	V	V	V	V	V
	Лимфоциты	0,73	V	V	V	V	V	
	Моноциты	7,19	V	V	V	V	V	V
	СОЭ (скорость оседания эритроцитов)	7,43	V	V	V	V	V	V
	Глюкоза	4,56				V	V	V
	Общий билирубин	7,33				V	V	V
	АЛТ (аланинаминотрансфераза)	17,75				V	V	V
	АСТ (аспартатаминотрансфераза)	3,90				V	V	V
	Общий белок	0,17				V	V	

По данным таблицы 1, комбинация № I включает в себя классические характеристики, на совокупности которых устанавливается диагноз туберкулеза [8].

Комбинация № II содержит показатели аналогичные комбинации I, но без учета жалоб, подозрительных на туберкулез [2, 8].

Комбинация № III содержит показатели, аналогичные комбинации № I, без уточнения факта наличия социальной группы риска, нахождения в МЛС, сопутствующей патологии [2, 4, 8].

Комбинация № IV сформирована с учетом мнения эксперта, но с включением минимального, на их взгляд, количества признаков: возрастно-половые характеристики, факт трудоустройства, принадлежность к фтизиатрической группе риска и показатели общего анализа крови. Такой вариант возможен в случаях отсутствия полной информации о пациенте.

Комбинация № V создана на основе мнения экспертов для экстремальных случаев, когда нет контакта с пациентом: возрастно-половые характеристики, рентгенологические симптомы и показатели общего анализа крови.

Комбинация № VI сформирована на основании симптомов, наиболее подозрительных на туберкулез, в сочетании с клиническими проявлениями иммунодефицитного состояния и показателей общего анализа крови.

РЕЗУЛЬТАТЫ АНАЛИЗА КОМБИНАЦИЙ ПАРАМЕТРОВ ПАЦИЕНТОВ

Для анализа всех групп признаков использовался следующий подход: в начале делалась оценка информативности признаков функцией f-classif, а затем решалась задача классификации классическими методами машинного обучения (МО).

При анализе информативности признаков, объединенных в комбинацию I (табл. 2), выявлено следующее.

Таблица 2 – Информативность признаков, объединенных в комбинацию I

Обозначение	f-classif
Фтизиатрическая группа риска	97,20816
RBC (эритроциты)	34,61265
WBC (лейкоциты)	11,4486
HGB (гемоглобин)	24,45369
Палочкоядерные нейтрофилы	5,309416
Сегментоядерные нейтрофилы	14,47767
Отсутствие флюорографии 2 и более лет	0,404151
Лимфоциты	19,84921
Моноциты	10,47143
СОЭ (скорость оседания эритроцитов)	0,738235
Поликлиническая группа риска	0,682746
Употребление алкоголя	145,5263
Употребление наркотиков	0,015476
Курение	0,015838
Выявление при обращении	8,433119
Выявление при профосмотре	42,11772
Факт установленного контакта	0,727153
Факт нахождения в местах лишения свободы	0,58887
Снижение массы тела	8,690183
Потливость	1,434859
Температура 37 и более	0,365746
Слабость при выполнении обычной нагрузки	8,456986
Кашель	12,46309
Кровохарканье	25,11028
Боли в груди	4,992426
Факт наличия сопутствующей патологии	2,118752
Возраст	31,59034
Пол	0,906338
Факт трудоустройства	320,5239

По данным таблицы 1 выявлена наибольшая классифицирующая функция у показателя, оценивающего факт трудоустройства (320,5), несколько меньше – факт употребления алкоголя (145,5) и принадлежность к социальной группе (97,8). Наименьшую классифицирующую функцию имели факт курения (0,01), повышение температуры выше 37 °С (0,36) и отсутствие флюорографии 2 и более лет (0,4).

Метрики качества различных методов МО по комбинации I приведены в таблице 3. Выявлено, что все методы имеют высокую аккуратность и точность в пределах 0,86–0,99.

Таблица 3 – Метрики качества методов для комбинации I

Методы МО	Аккуратность	Точность	Площадь под ROC-кривой
Линейное разделение	0,918919	0,992126	0,914399
Опорных векторов	0,932432	0,996063	0,94424
Логистическая регрессия	0,935811	0,992126	0,931364
Ближайших соседей	0,905405	0,980315	0,853554
Наивный Байес	0,847973	0,862205	0,716972
Случайный лес	0,935811	0,988189	0,918313
Дерево решений	0,902027	0,952756	0,805019

По данным таблицы, площадь ROC под кривой колеблется от 71,6 % (при методе «наивный Байес») до 94,4 % (при методе «опорных векторов»).

При оценке информативности признаков объединенных в комбинацию II (табл. 4) выявлено наличие классифицирующей функции в пределах 0,4–97,2.

Таблица 4 – Информативность признаков, объединенных в комбинацию II

Обозначение	f-classif
Фтизиатрическая группа риска	97,20816
WBC (лейкоциты)	11,4486
HGB (гемоглобин)	24,45369
Палочкоядерные нейтрофилы	5,309416
Сегментоядерные нейтрофилы	14,47767
Отсутствие флюорографии 2 и более лет	0,404151
Лимфоциты	19,84921
Моноциты	10,47143
СОЭ (скорость оседания эритроцитов)	0,738235
Факт выявления при обращении	8,433119
Факт выявления при профосмотре	42,11772
Факт выявления сопутствующей патологии	2,118752
Возраст	31,59034

При математической интерпретации модели, составленной на основе показателей, объединенных в комбинацию II, выявлено, что все методы имеют высокую аккуратность и точность в пределах 0,84–1,00 (табл. 5).

Таблица 5 – Метрики качества методов для комбинации II

Методы МО	Аккуратность	Точность	Площадь под ROC-кривой
Линейное разделение	0,89527027	0,980314961	0,833679654
Опорных векторов	0,87162162	1	0,934931507
Логистическая регрессия	0,885135135	0,980314961	0,808952838
Ближайших соседей	0,841216216	0,948818898	0,628658009
Наивный Байес	0,810810811	0,858267717	0,647638366
Случайный лес	0,841216216	0,92519685	0,658889273
Дерево решений	0,851351351	0,917322835	0,692578125

Согласно таблице 4, выявлено, что методы показывают преимущественно площадь ROC под кривой от 0,628 (62,8 %) (метод «ближайших соседей») до 0,692 (69,2 %) (метод «дерево решений»), только при методе «опорных векторов» достигает 0,934 (93,4 %). Данный факт делает модель менее применимой для определения принадлежности пациентов в группу больных туберкулезом или исключения из нее.

При статистическом анализе показателей, объединенных в комбинацию III (табл. 6), выявлено наличие классифицирующей функции в пределах 0,7–97,2, что несколько меньше, чем при анализе комбинации I.

Таблица 6 – Информативность признаков, объединенных в комбинацию III

Обозначение	f-classif
Фтизиатрическая группа риска	97,20816
WBC (лейкоциты)	11,4486
HGB (гемоглобин)	24,45369
Палочкоядерные нейтрофилы	5,309416
Сегментоядерные нейтрофилы	14,47767
Отсутствие флюорографии 2 и более лет	0,404151
Лимфоциты	19,84921
Моноциты	10,47143
СОЭ (скорость оседания эритроцитов)	0,738235
Выявление при обращении	8,433119
Выявление при профосмотре	42,11772
Факт наличия сопутствующей патологии	2,118752
Возраст	31,59034

Дальнейшая математическая интерпретация модели, составленной на основе показателей, объединенных в комбинацию III, показала высокую аккуратность и точность в пределах 0,81–1,00 (табл. 7).

Таблица 7 – Метрики методов для комбинации III

Методы МО	Аккуратность	Точность	Площадь_под_ROC_кривой
Линейное разделение	0,89527027	0,980314961	0,833679654
Опорных векторов	0,871621622	1	0,934931507
Логистическая регрессия	0,885135135	0,980314961	0,808952838
Ближайших соседей	0,841216216	0,948818898	0,628658009
Наивный Байес	0,810810811	0,858267717	0,647638366
Случайный лес	0,841216216	0,92519685	0,658889273
Дерево решений	0,851351351	0,917322835	0,692578125

По данным таблицы 6, площадь под ROC-кривой находилась, по данным разных методов, от 0,62 (метод «ближайших соседей») до 0,93 (метод «опорных векторов»), что допускало оптимальный прогноз модели, построенной на основе комбинации 3.

Статистический анализ показателей, объединенных в комбинацию IV (табл. 8), выявил наличие классифицирующей функции в пределах 0,7–0,346, что существенно отличалось от других комбинаций и снижало вероятность использования этой комбинации в качестве диагностической модели.

Таблица 8 – Информативность признаков, объединенных в комбинацию IV

Обозначение	f-classif
RBC (эритроциты)	34,61265
WBC (лейкоциты)	11,4486
PLT (тромбоциты)	0,606842
HGB (гемоглобин)	24,45369
эозинофилы	3,586745
Палочкоядерные нейтрофилы	5,309416
Сегментоядерные нейтрофилы	14,47767
Лимфоциты	19,84921
Моноциты	10,47143
СОЭ (скорость оседания эритроцитов)	0,738235
Глюкоза	7,198143
Общий билирубин	7,43185
АЛТ (аланинаминотрансфераза)	4,569071
АСТ (аспартатаминотрансфераза)	7,338348
Общий белок	17,75764
Возраст	31,59034
Пол	0,906338

Дальнейшая математическая интерпретация модели, составленной на основе показателей, объединенных в комбинацию III, показала высокую аккуратность и точность в пределах 0,81–1,00 (табл. 9).

Таблица 9 – Метрики качества методов для комбинации IV

Методы МО	Аккуратность	Точность	Площадь под ROC-кривой
Линейное разделение	0,851351	0,988189	0,554795
Опорных векторов	0,861486	1	0,930508
Логистическая регрессия	0,85473	0,992126	0,596701
Ближайших соседей	0,844595	0,96063	0,623214
Наивный Байес	0,692568	0,692913	0,601122
Случайный лес	0,851351	0,952756	0,668879
Дерево решений	0,736486	0,834646	0,489919

По данным таблицы 8, площадь под ROC-кривой находится преимущественно в пределах от 0,48 (метод «дерево решений») до 0,66 (66 %) (метод «случайный лес») и только методом «опорных векторов» достигает 0,93, что существенно ниже, чем при анализе других комбинаций и порогового значения 50 %. Модель, построенная на комбинации показателей, представленных в таблице 7, не позволяет правильно определить принадлежность пациента к группе больных туберкулезом.

Статистический анализ показателей, объединенных в комбинацию V (табл. 10), выявил наличие классифицирующей функции в пределах 0,003–24,4, что исключало модель, построенную из комбинаций показателей, из ряда перспективных.

Таблица 10 – Информативность признаков, объединенных в комбинацию V

Обозначение	f-classif
RBC (эритроциты)	34,61265
WBC (лейкоциты)	11,4486
PLT (тромбоциты)	0,606842
HGB (гемоглобин)	24,45369
эозинофилы	3,586745
Палочкоядерные нейтрофилы	5,309416
Сегментоядерные нейтрофилы	14,47767
Лимфоциты	19,84921
Моноциты	10,47143
СОЭ (скорость оседания эритроцитов)	0,738235
Глюкоза	7,198143
Общий билирубин	7,43185
АЛТ (аланинаминотрансфераза)	4,569071
АСТ (аспартатаминотрансфераза)	7,338348
Общий белок	17,75764
Возраст	31,59034
Пол	0,906338
Рентгенологическое определение очагового образования / нескольких очагов	0,003174
Деструкции	0,289893
Утолщение плевры/спайки	20,45498
Кальцинаты, выявленные при лучевых методах	0,121052
Усиленный сосудистый рисунок	6,721637

Этот вывод подтверждался низкой аккуратностью (16 % при использовании метода «опорных векторов») и невысокой точностью (3 %) при математической обработке модели, составленной из показателей, объединенных в комбинацию V (табл. 11).

Таблица 11 – Метрики качества методов для комбинации V

Методы МО	Аккуратность	Точность	Площадь под ROC-кривой
Линейное разделение	0,864864865	0,976377953	0,725430598
Опорных векторов	0,168918919	0,031496063	0,572916667
Логистическая регрессия	0,858108108	0,976377953	0,686619718
Ближайших соседей	0,837837838	0,960629921	0,575481256
Наивный Байес	0,689189189	0,677165354	0,61287835
Случайный лес	0,824324324	0,929133858	0,590883191
Дерево решений	0,733108108	0,795275591	0,552988333

По данным таблицы 10, площадь под ROC-кривой находится в пределах от 0,55 (метод «дерево решений») до 0,72 (метод «линейное разделение»), что наблюдается в этой комбинации. Модель,

построенная на данной комбинации, не может определять принадлежность пациента к группе больных туберкулезом и/или другим заболеваниям органов дыхания.

Анализ признаков, объединенных в комбинацию VI (табл. 12), выявил разброс классифицирующих функций в пределах 0,01–320,5, что свидетельствовало о неоднородности выбранных показателей и малой вероятности построения оптимальной модели.

Таблица 12 – Информативность признаков, объединенных в комбинацию VI

Обозначение	f-classif
Фтизиатрическая группа риска	97,20816
RBC (эритроциты)	34,61265
WBC (лейкоциты)	11,4486
HGB (гемоглобин)	24,45369
Сегментоядерные нейтрофилы	14,47767
Лимфоциты	19,84921
Моноциты	10,47143
Общий белок	17,75764
Социальная группа риска	20,45498
Кашель	12,46309
Патология ЛОР-органов	10,17111
Возраст	31,59034
Факт перенесенной пневмонии	32,60137
Онкологические заболевания	129,7235
Патология сердечно-сосудистой системы	30,12429
Факт трудоустройства	320,5239

По данным таблицы 12 видно, что комбинация VI включает наибольшее множество переменных, чьи показатели классифицирующей функции находятся в пределах 10,0–320,5. Это характеризует неоднородность выборки, и модель, построенная на ней, не может быть рассмотрена в дальнейшем и подтверждается низкой точностью (0,90–0,92) предсказания и аккуратностью (0,59–0,83) (табл. 13).

Таблица 13 – Метрики качества методов для комбинации VI

Методы МО	Аккуратность	Точность	Площадь под ROC-кривой
Линейное разделение	0,942568	0,980315	0,905583
Опорных векторов	0,952703	0,984252	0,925214
Логистическая регрессия	0,942568	0,980315	0,905583
Ближайших соседей	0,89527	0,980315	0,83368
Наивный Байес	0,587838	0,543307	0,597588
Случайный лес	0,918919	0,96063	0,841289
Дерево решений	0,898649	0,92126	0,787201

По данным таблицы 13, наименьший показатель площади под ROC-кривой показал метод «Наивный Байес» (58 %), что недостаточно для признания модели, построенной на комбинации показателей, оптимальной.

ЗАКЛЮЧЕНИЕ

Таким образом, комбинации имели следующие площади под ROC-кривой:

– комбинация I: 0,71–0,94;
– комбинация II: 0,61–0,93;
– комбинация III: 0,62–0,93;
– комбинация IV: 0,48–0,93;
– комбинация V: 0,57–0,72;
– комбинация VI: 0,59–0,92.

Наиболее качественные во всех комбинациях прогнозы показывал метод «логистическая регрессия». Полученный расчет площади под ROC-кривой показал перспективность комбинации I, построенной на основании классических признаков с включением: возрастно-половых показателей; жалоб, подозрительных на туберкулез; наличия «фтизиатрической» (наблюдение в IV, VI диспансерных группах), социальной группы риска (постояльцы и работники учреждений «закрытого» типа, лица без определенного места жительства и/или находившиеся в системе исполнения наказаний и т. д.); факта наличия трудоустройства; отсутствие флюорографии 2 и более лет; медицинской

группы риска (наличие хронической патологии, в том числе с иммунокопрометацией); вредных привычек (употребление алкоголя, курение); метода выявления, факта пребывания в местах лишения свободы, а также показателей общего анализа крови.

На основании расчетов, приведенных в данной статье, был составлен план дальнейших исследований:

1) определить наиболее информативные признаки в комбинации I;
2) оценить перспективность построения РП на основе ансамблей линейных разделителей;
3) минимизировать число признаков только на основе математической модели [9], т. е. исключить влияние человеческого фактора;
4) сравнить результаты человеко-машинной процедуры поиска РП с результатами математической модели минимизации числа входных признаков.

Список источников

1. Алексеева, Г. И. Бактерионосительство или латентный туберкулез? / Г. И. Алексеева, А. Ф. Кравченко // Acta Biomedica Scientifica. – 2012. – № 5–1 (87).

2. Клинические рекомендации «Туберкулез у взрослых». – 2022. – URL: https://cr.minzdrav.gov.ru /schema/16_2.

3. Отраслевые и экономические показатели противотуберкулёзной работы в 2019–2020 гг. Аналитический обзор основных показателей и статистические материалы / под ред. С. А. Стерликова. – Москва : РИО ЦНИИОИЗ, 2021. – 63 с.

4. Приказ Министерства здравоохранения РФ от 15 ноября 2012 г. № 932н «Об утверждении Порядка оказания медицинской помощи больным туберкулезом». – URL: https://cr.minzdrav.gov.ru/documents/9119.

5. Синицына, А. В. Оценка эффективности различных методов лучевой диагностики в выявлении туберкулёза у детей / А. В. Синицына, П. В. Гаврилов, А. В. Синицын, С. В. Михайлова, К. В. Прибыток, Е. В. Синельникова // Педиатр. – 2017. – Т. 8, вып. 3. – С. 94–100.

6. Старшинова А. А. Иммунодиагностика туберкулёза сегодня: современные иммунологические тесты и дифференцированный подход к их применению в практике / А. А. Старшинова, И. Ф. Довгалюк, Ю. С. Зинченко, Д. А. Кудлай, П. К. Яблонский // Практическая пульмонология. – 2019. – № 2. – С. 28–32.

7. Тюлькова, Т. Е. Практическое применение методов машинного обучения на примере определения активности туберкулезного процесса у лиц с минимальными туберкулезными изменениями, выявленными не рентгенограмме органов грудной клетки / Т. Е. Тюлькова, П. Ф. Чернавин, Н. П. Чернавин // Клинический вестник ФМБЦ им. А.И. Бурназяна. – 2022. – № 2. – С. 64–73.

8. Фтизиатрия: национальное руководство / гл. ред. М. И. Перельман. – Москва : ГЭОТАР-Медиа, 2007. – 505 с.

9. Чернавин, П. Ф. Управление качеством решающего правила и минимизация числа признаков в задачах классификации на основе моделей математического программирования / П. Ф. Чернавин, Н. П. Чернавин, Ф. П. Чернавин // Прикаспийский журнал: управление и высокие технологии. – 2023. – № 1. – С. 112–119.

10. Datta, S. Comparison of sputum collection methods for tuberculosis diagnosis: a systematic review and pairwise and network meta-analysis / S. Datta, L. Shah, R. H. Gilman, C. A. Evans // Lancet Glob Health. – 2017. – № 5 (8). – P. e760–e771.

11. Divala, T. H. Accuracy and consequences of using trialof-antibiotics for TB diagnosis (ACT-TB study): protocol for a randomised controlled clinical trial / T. H. Divala, K. L. Fielding, D. J. Sloan, et al. // BMJ Open 2020. – № 10. – P. e033999. – doi: 10.1136/ bmjopen-2019-033999.

12. Zuberi, F. F. Role of Bronchial Washing Gene Xpert in Sputum-Scarce Cases of Suspected Pulmonary Tuberculosis / F. F. Zuberi, S. Hussain, S. Hameed, B. F. Zuberi // Pak. J. Med. Sci. – 2019. – № 35 (1). – P. 211–214.

References

1. Alekseeva, G. I., Kravchenko, A. F. Bacteriocarrier or latent tuberculosis? *Acta Biomedica Scientifica*, 2012, no. 5–1 (87).

2. *Clinical guidelines "Tuberculosis in adults"*, 2022. URL: https://cr.minzdrav.gov.ru/schema/16_2.

3. Sterlikov, S. A. (ed.) *Industry and economic indicators of TB work in 2019–2020. Analytical review of key indicators and statistical materials*. Moscow, RIO TsNIIOIZ, 2021. 63 p.

4. *Order of the Ministry of Health of the Russian Federation of November 15, 2012 № 932 "On Approval of the Procedure for Providing Medical Care to Patients with Tuberculosis".*

5. Sinitsyna, A. V., Gavrilov, P. V., Sinitsyn, A. V., Mikhailova, S. V., Pribitok, K. V., Sinelnikova, E. V. Evaluation of efficiency of different methods of radiation diagnosis in the detection of tuberculosis in children. *Pediatrician*, 2017, vol. 8, no. 3, pp. 94–100.

6. Starshinova, A. A., Dovgalyuk, I. F., Zinchenko, Yu. S., Kudlay, D. A., Yablonsky, P. K. Immunodiagnosis of tuberculosis: modern immunological tests and differentiated approach to their use in practice. *Practical pulmonology*, 2019, no. 2, pp. 28–32.

7. Tyulkova, T. E. Chernavin, P. F., Chernavin, N. P. Practical application of machine learning methods on the example of determining the activity of the tuberculosis process in individuals with minimal tuberculous changes detected on chest X-ray. *Clinical Bulletin of the FMBC named after A.I. Burnazyan*, 2022, no. 2, pp. 64–73.

8. Perelman, M. I. (editor-in-chief) *Phthisiology: national leadership.* Moscow, GEOTAR-Media Publ., 2007.

9. Chernavin, P. F., Chernavin, N. P., Chernavin, F. P. Control of the quality of the decision rule and minimization of the number of features in classification problems based on mathematical programming models. *Caspian Journal: Control and High Technologies*, 2023, no. 1, pp. 112–119.

10. Datta, S., Shah, L., Gilman, R. H., Evans, C. A. Comparison of sputum collection methods for tuberculosis diagnosis: a systematic review and pairwise and network meta-analysis. *Lancet Glob Health*, 2017, no. 5 (8), pp. 760–771.

11. Divala, T. H., Fielding, K. L., Sloan, D. J. et al. Accuracy and consequences of using trialof-antibiotics for TB diagnosis (ACT-TB study): protocol for a randomised controlled clinical trial. *BMJ Open*, 2020, no. 10, p. e033999. doi: 10.1136/ bmjopen-2019-033999.

12. Zuberi, F. F, Hussain, S., Hameed, S., Zuberi, B. F. Role of Bronchial Washing Gene Xpert in Sputum-Scarce Cases of Suspected Pulmonary Tuberculosis. *Pak. J. Med. Sci.*, 2019, no. 35 (1), pp. 211–214.

Статья поступила в редакцию 25.07.2023; одобрена после рецензирования 31.08.2023; принята к публикации 08.09.2023.

The article was submitted 25.07.2023; approved after reviewing 31.08.2023; accepted for publication 08.09.2023.

УДК 004.001

ИСПОЛЬЗОВАНИЕ АЛГОРИТМА МАШИННОГО ОБУЧЕНИЯ НЕЙРОННОЙ СЕТИ ДЛЯ РЕШЕНИЯ ПРОБЛЕМ БЕЗОПАСНОСТИ КОМПЬЮТЕРНЫХ СЕТЕЙ

Путято Михаил Михайлович, Кубанский государственный технологический университет, 350072, Российская Федерация, г. Краснодар, ул. Московская, 2,
кандидат технических наук, доцент, ORCID: 0000-0003-0414-6034, e-mail: putyato.m@gmail.com
Макарян Александр Самвелович, Кубанский государственный технологический университет, 350072, Российская Федерация, г. Краснодар, ул. Московская, 2,
кандидат технических наук, доцент, ORCID: 0000-0002-1801-6137, e-mail: msanya@yandex.ru
Черкасов Александр Николаевич, Кубанский государственный технологический университет, 350072, Российская Федерация, г. Краснодар, ул. Московская, 2,
кандидат технических наук, доцент, ORCID: 0000-0002-5015-4556, e-mail: cherk@mail.ru
Левченко Алина Михайловна, Кубанский государственный технологический университет, 350072, Российская Федерация, г. Краснодар, ул. Московская, 2,
ассистент, ORCID: 0009-0003-3706-9992, e-mail: alikolcova@yandex.ru

На данный момент машинное обучение является одним из актуальных направлений в сфере развития информационных технологий. Представляя собой определенный анализ каких-либо данных, машинное обучение при постановке определенной задачи и разработанного алгоритма может быть направлено на создание программного продукта. В данной статье был использован алгоритм машинного обучения нейронной сети на примере датасет «UNSW-NB15». Используя математические методы в огромных наборах данных алгоритмы машинного обучения, по сути, строят модели поведения и используют эти модели в качестве основы для будущих прогнозов на основе новых входных данных. Применяя машинное обучение, специалисты по информационной безопасности решают вопросы защиты компьютерной сети, актуальные для предприятий. В статье приводится алгоритм отбора пакета данных для машинного обучения, также проведен анализ его использования. Далее были описаны алгоритмы для проведения 2 тестов: с расширением и ручной подгонкой. В результате их проведения были получены кривые, которые демонстрируют меру производительности используемых алгоритмов. Полученные данные позволяют судить о преимуществах использования некоторых алгоритмов методов машинного обучения с использованием нейронных сетей.

Ключевые слова: машинное обучение, нейронные сети, алгоритм, компьютерные сети, пакет данных

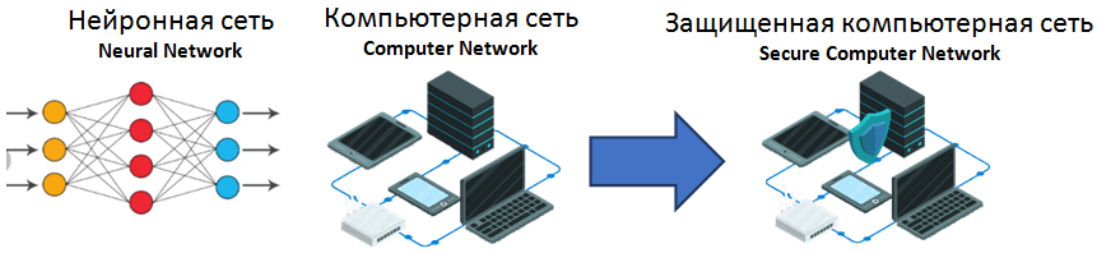
USING THE NEURAL NETWORK MACHINE LEARNING ALGORITHM TO SOLVE COMPUTER NETWORK SECURITY PROBLEMS

Putyato Michael M., Kuban State Technological University, 2 Moskovskaya St., Krasnodar, 350072, Russian Federation,
Cand. Sci. (Engineering), Associate Professor, ORCID: 0000-0001-9974-7144, e-mail: putyato.m @gmail.com
Makaryan Alexander S., Kuban State Technological University, 2 Moskovskaya St., Krasnodar, 350072, Russian Federation,
Cand. Sci. (Engineering), Associate Professor, ORCID: 0000-0002-1801-6137, e-mail: msanya@yandex.ru
Cherkasov Alexander N., Kuban State Technological University, 2 Moskovskaya St., Krasnodar, 350072, Russian Federation,
Cand. Sci. (Engineering), Associate Professor, ORCID: 0000-0002-5015-4556, e-mail: cherk@mail.ru
Levchenko Alina M., Kuban State Technological University, 2 Moskovskaya St., Krasnodar, 350072, Russian Federation,
assistant, ORCID: 0009-0003-3706-9992, e-mail: alikolcova@yandex.ru

At the moment, machine learning is one of the most relevant areas in the field of information technology development. Representing a certain analysis of any data, machine learning, when setting a certain task and a developed algorithm, can be aimed at creating a software product. In this article, a neural network machine learning algorithm was applied using the example of the «UNSW-NB15» dataset. Using mathematical methods in huge datasets, machine learning algorithms, in fact, build behavior models and use these models as a basis for future predictions based on new input data Using machine learning, information security specialists solve computer network protection issues relevant to enterprises. The article provides an algorithm for selecting a data package for machine learning, as well as an analysis of its use. Next, algorithms for conducting 2 tests were described: with expansion and manual adjustment. The data obtained allow us to judge the advantages of using some algorithms of machine learning methods using neural networks.

Keywords: machine learning, neural networks, algorithm, computer networks, data package

Graphic annotation (Графическая аннотация)



ВВЕДЕНИЕ

Вопрос сетевой безопасности сегодня как никогда важен для любой организации. Количество и сложность угроз сетевой безопасности растет с каждым днем. Абсолютно любая информация, обладающая какой-либо ценностью (финансовая, политическая, конкурентная, военная), может быть подвергнута угрозе кибератаки [1].

Обратимся к статистике заражения вредоносным ПО, распределенным по годам, показанной на рисунке 1. К такому программному обеспечению можно отнести: трояны, программы-шпионы, программы-вымогатели, рекламное ПО, черви и т. п. Можно заметить огромную тенденцию роста на рисунке 1.

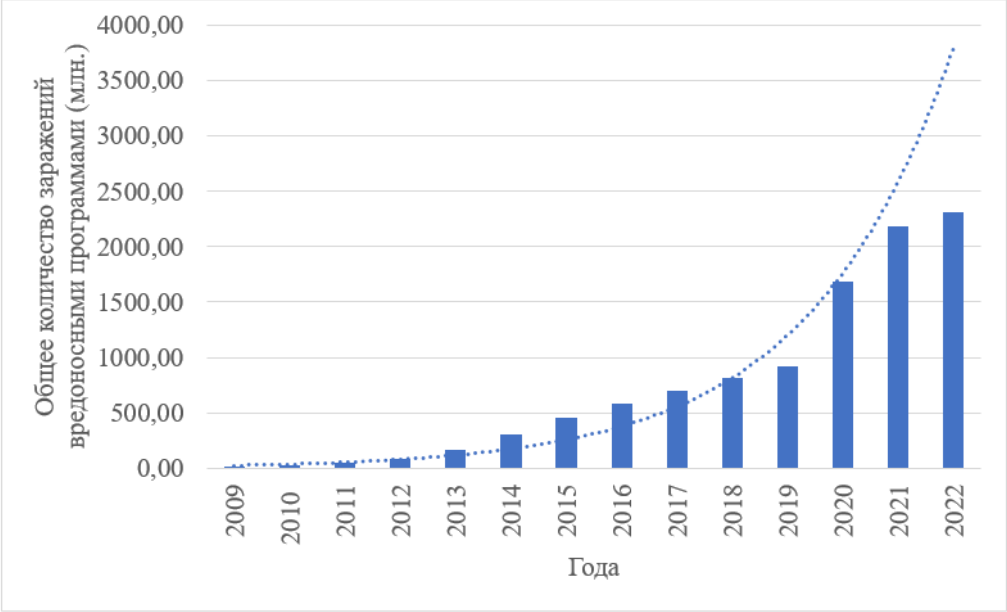


Рисунок 1 – Статистика заражения вредоносным ПО, распределенная по годам [1]

Для бизнеса сегодня наибольшую угрозу представляют атаки, связанные с утечкой данных, поскольку любому предприятию необходимо хранить информацию, недоступную для ее конкурентов [2].

Современные подходы для решения вопросов информационной безопасности основаны на нейронных сетях. Используя математические методы в огромных наборах данных, алгоритмы машинного обучения, по сути, строят модели поведения и используют эти модели в качестве основы для будущих прогнозов на основе новых входных данных[6]. Следует отметить, что применение методов машинного обучения уже сегодня является популярной практикой для решения задач безопасности, и с каждым годом доля рынка таких программных продуктов будет только расти.

Поставленная в данной статье задача – применение методов машинного обучения нейронной сети для решения проблем безопасности компьютерных сетей.

МЕТОДЫ И МАТЕРИАЛЫ ИССЛЕДОВАНИЯ

Более правдивым набором данных для исследования является UNSW-NB15. Набор данных безопасности компьютерных сетей UNSW-NB15 был выпущен в 2015 г. (Moustafa & Slay, 2015). Этот набор данных состоит из 2 540 044 реалистичных современных записей нормальной и аномальной сетевой активности. Эти записи были собраны генератором трафика IXIA с использованием трех

виртуальных серверов. Два сервера были настроены для распределения обычного сетевого трафика, а третий был настроен на создание аномального сетевого трафика.

Всего было извлечено 49 параметров пакетов, обработанных инструментами Argus и Bro-IDS и характеризуемых определенными функциями.

Функции на основе пакетов извлекаются из заголовка пакета и его полезной нагрузки (называемой пакетными данными). Функции, основанные на потоке, генерируются с помощью упорядочивания пакетов от источника к получателю, перемещающихся по сети.

Направление, длина между пакетами и время между приходами являются наиболее важными свойствами в формулировке характеристик на основе потока: общая продолжительность (dur) и время жизни от места назначения до источника (dttl).

Функции делятся на три группы, а именно:

а) основные (от 6 до 18);

б) содержание (от 19 до 26);

в) время (от 27 до 35).

Элементы с 36 по 40 и с 41 по 47 помечены как функции общего назначения и функции подключения соответственно. Категория функций общего назначения включает в себя те функции, которые предназначены для определения цели отдельной записи, в то время как признаки соединения изображают характеристику соединения среди ста последовательно отправленных пакетов.

Последние две функции включают категории атак и метки. Атаки подразделяются на анализирующие атаки, бэкдоры, DoS, эксплойты, фаззеры, атаки общего назначения, шелл-коды и черви [7]. Нормальные сетевые пакеты представлены 2 218 761 записью, в то время как фаззеры, анализирующие атаки, бэкдоры, DoS, эксплойты, атаки общего назначения, шелл-коды и сигнатуры червей, включают 24246, 2677, 2329, 16535, 44525, 215481, 13987, 1511 и 174 записей соответственно.

Следовательно, существует значительный недостаток баланса для набора данных, поскольку 87 % набора данных состоит из обычных записей. Разработчики набора данных также провели подвыборку и разделили набор данных на данные для обучения и тестирования, представленные в таблице 1.

Таблица 1 – Характеристики наборов данных

Класс записи	**Кол-во в данных для обучения**	**Кол-во в данных для проверки**
Нормальные пакеты	56000	37000
Анализирующие атаки	2000	677
Бэкдоры	1746	583
DoS	12264	4089
Эксплоиты	33393	11132
Фаззеры	18184	6062
Атаки общего назначения	40000	18871
Разведывательные атаки	10491	3496
Шелл-коды	1333	378
Черви	130	44
Общее количество	175341	82332

Структура этого набора данных более сложная по сравнению с другими эталонными наборами данных, такими как DARPA98, KDDCUP 99 и NSL-KDD. Это делает UNSW-NB15 более полным для более надежной оценки существующих систем обнаружения сетевых вторжений.

В проведенном исследовании «Analysis of KDD-Cup’99, NSL-KDD and UNSW-NB15 Data setsusing Deep Learningin Io T» [5], которое использует этот же набор данных, пришли к выводу, что достаточно взять лишь некоторые характеристики данных для построения достаточно точной нейронной сети, которые описаны в таблице 2.

Таблица 2 – Выбранные данные для обучения сети в исследовании «Analysis of KDD-Cup’99, NSL-KDD and UNSW-NB15»

Название	**Описание**
dur	Общее время записи
service	Сервис (http, ftp, smtp, ssh, dns, ftp-data ,irc)
dpkts	Кол-во пакетов, отправленных на сервер
sbytes	Кол-во байт, отправленных от клиента на сервер
dbytes	Кол-во байт, отправленных от сервера на клиент
sloss	Флаг принятия пакетов клиентом
dloss	Флаг принятия пакетов сервером
sinpkt	Время прибытия обратного пакекта

В этом исследовании использовалась глубокая нейронная сеть (DNN) с 20 скрытыми слоями. 70 % набора данных было использовано для обучения, оставшиеся 30% были поровну разделены на наборы тестов и перекрестной проверки [8].

В данном случае использовать глубокую нейронную сеть не представляется возможным, так как необходимо производить вычисления в реальном времени. Для этого список характеристик был дополнен следующими параметрами (табл. 3).

Таблица 3 – Дополнение к выбранным характеристикам для исследования

Название	Характеристика
proto	Протокол передачи
sload	Кол-во битов в секунду, переданных от клиента
dload	Кол-во битов в секунду, переданных от сервера
sinpkts	Межпакетный интервал для клиента
dinpkts	Межпакетный интервал для сервера

Кроме того, для упрощения и ускорения работы нейросети, поставим задачу бинарной классификации (определение нормальный или аномальный пакет).

Для нормализации данных для обучения также следует определиться с тем, какие параметры будут категоризированы. К ним относятся: proto, service, sloss, dloss.

В итоге получен парсер, написанный на языке программирования C#, который конвертирует исходный файл к виду необходимой модели. Примеры строк исходного и результирующего файла изображены на рисунках 2 и 3 соответственно.

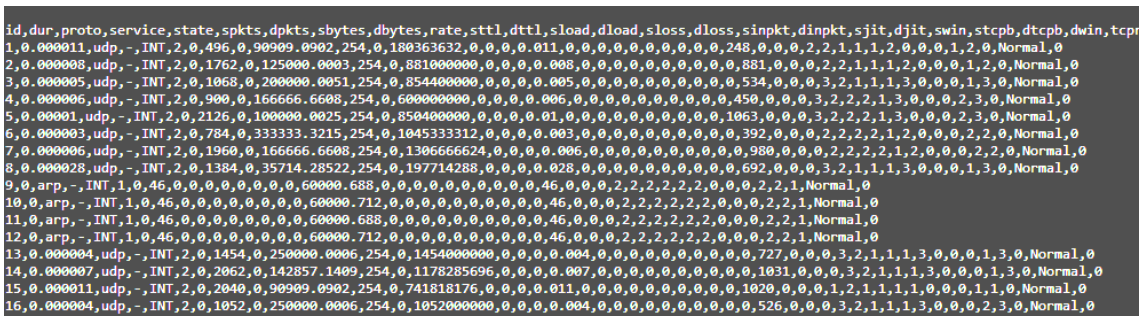


Рисунок 2 – Исходные данные до обработки

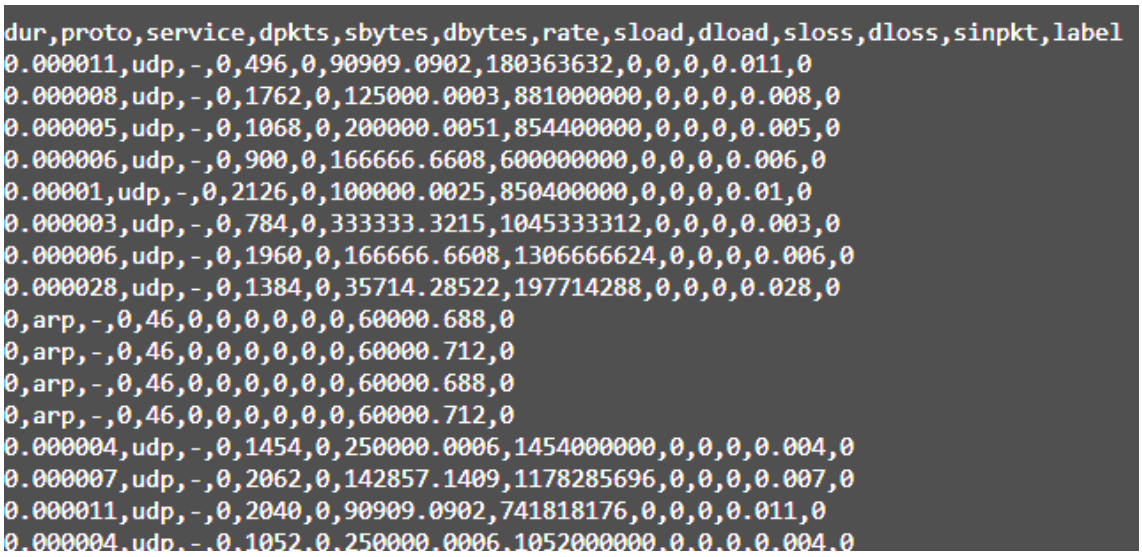


Рисунок 3 – Исходные данные после обработки

В качестве исходных данных были приняты несколько параметров с нормальным распределением, и 4 категоризированных параметра. Так как подавляющее большинство работы использует язык программирования C#, рассмотрим, какие варианты алгоритмов бинарной классификации предоставляет библиотека с открытым исходным кодом, разработанная Microsoft для платформы .net «ML.NET».

В данном случае будут рассмотрены только линейные алгоритмы, поскольку только они могут предоставить необходимую скорость принятия решения:

– стохастический двойной покоординатный подъем. Этот алгоритм основан на методе стохастического двухкоординатного восхождения (SDCA), современной методике оптимизации для выпуклых целевых функций. Именно его использовали в статье KDD [4];

– L-BFGS. Реализованный метод оптимизации основан на методе Бройдена – Флетчера – Гольдфарба–Шанно [5] с ограниченной памятью (L-BFGS);

– посимвольный стохастический градиентный спуск.

Также стоит отдельно выделить алгоритмы дерева решения. Такие алгоритмы в своей работе используют своеобразную блок-схему, которая содержит ряд решения функции. Для использования таких алгоритмов не надо как-либо нормализовать модель, потому что каждое отдельное значение в векторе признаков используется в процессе принятия решения независимо. Такие алгоритмы, как правило, очень точны.

К таким алгоритмам относятся:

– быстрое дерево, также называемое дерево принятия решений с градиентным бустингом. Подробно будет рассмотрено ниже;

– LightGbm – другая разновидность дерева решения, использующая своеобразную функцию для получения коэффициента корректировки;

– быстрый лес – алгоритм, в котором применяется множество деревьев, таких как в первом алгоритме из этой группы.

Было принято решение провести 2 теста: первый тест с использованием расширения для ML.NET под названием «ModelBuilder», а также более точный, уже с помощью ручной подгонки.

Для первого теста необходимо создать новый проект в среде разработки «VisualStudio» и добавить специальный шаблон под названием «Modelbuilder» (рис. 4).

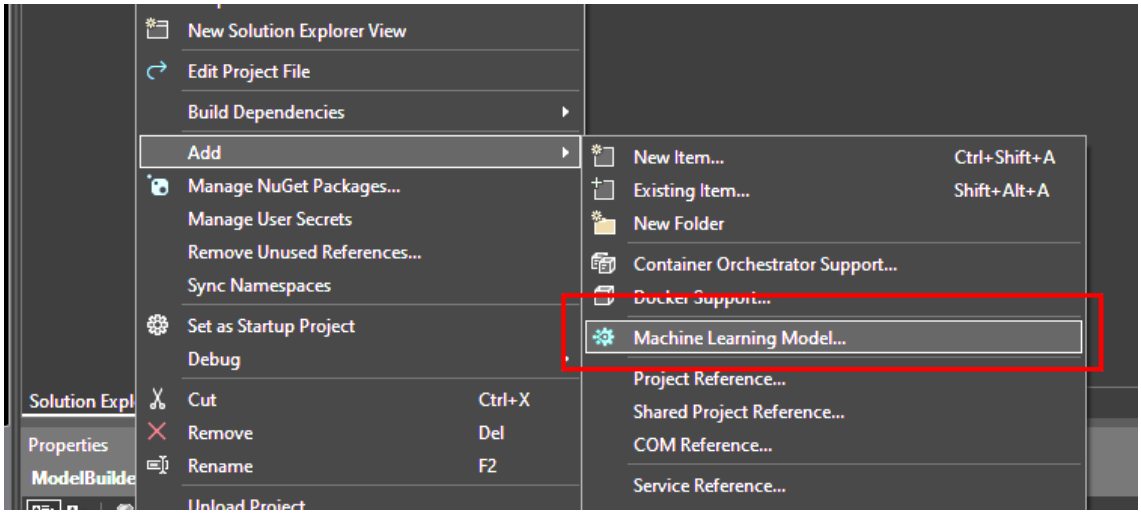


Рисунок 4 – Шаблон «Modelbuilder»

Следующим шагом выбираем сценарий, соответствующий нашей задачи, а именно «Dataclassification» (рис. 5).

Во вкладке Data (рис. 6) необходимо выбрать файл, который сгенерировал наш парсер. После его загрузки необходимо указать верные данные во вкладке «Advanceddataoptions», а именно соответствие данных правильным типам и указатели на то, что поле категоризировано (рис. 7).

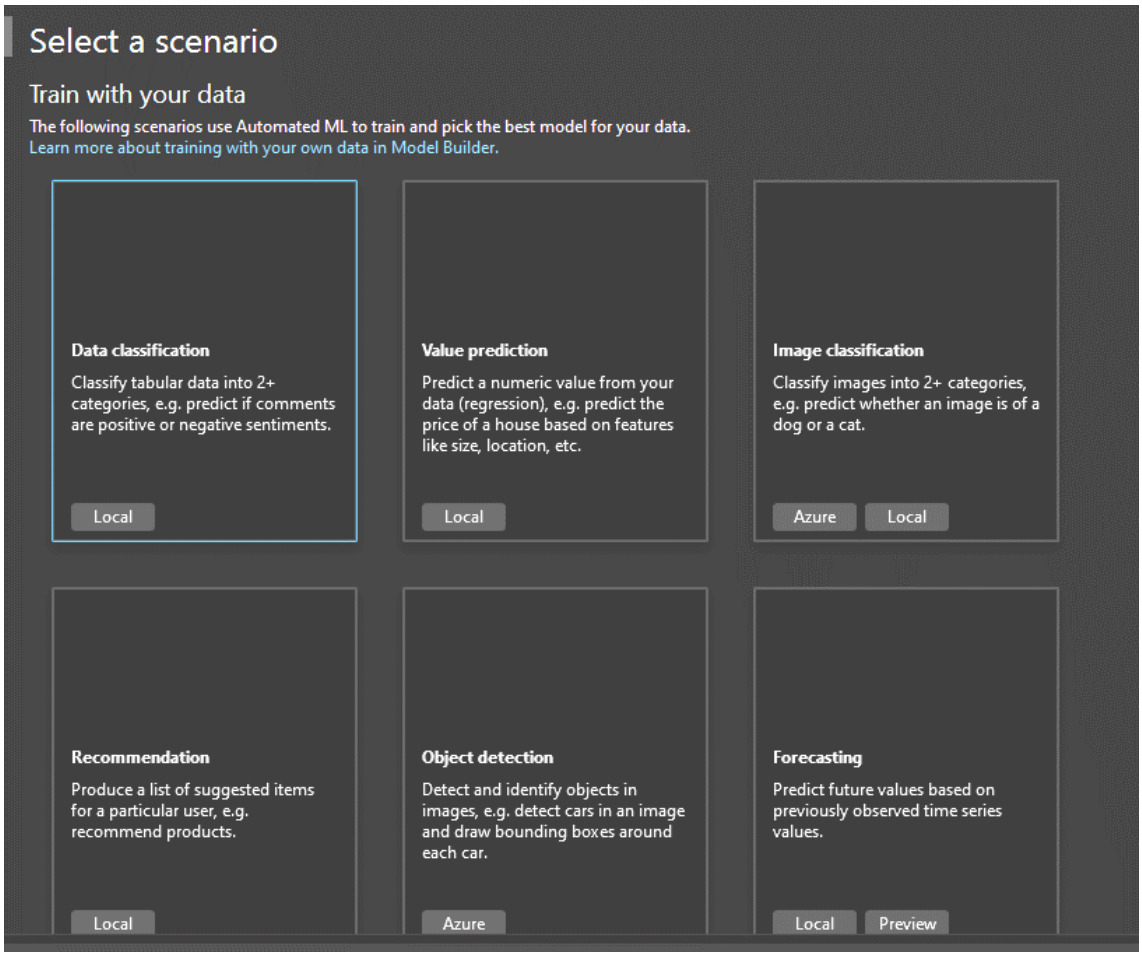


Рисунок 5 – Вкладка «Dataclassification»

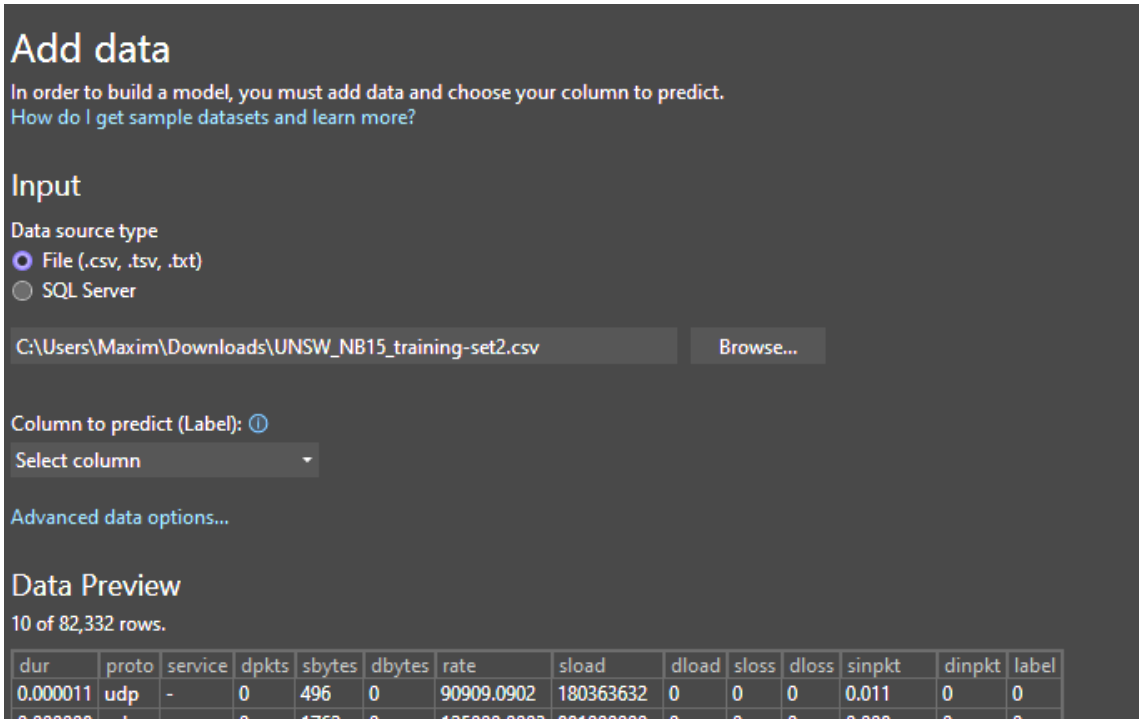


Рисунок 6 – Вкладка «Data»

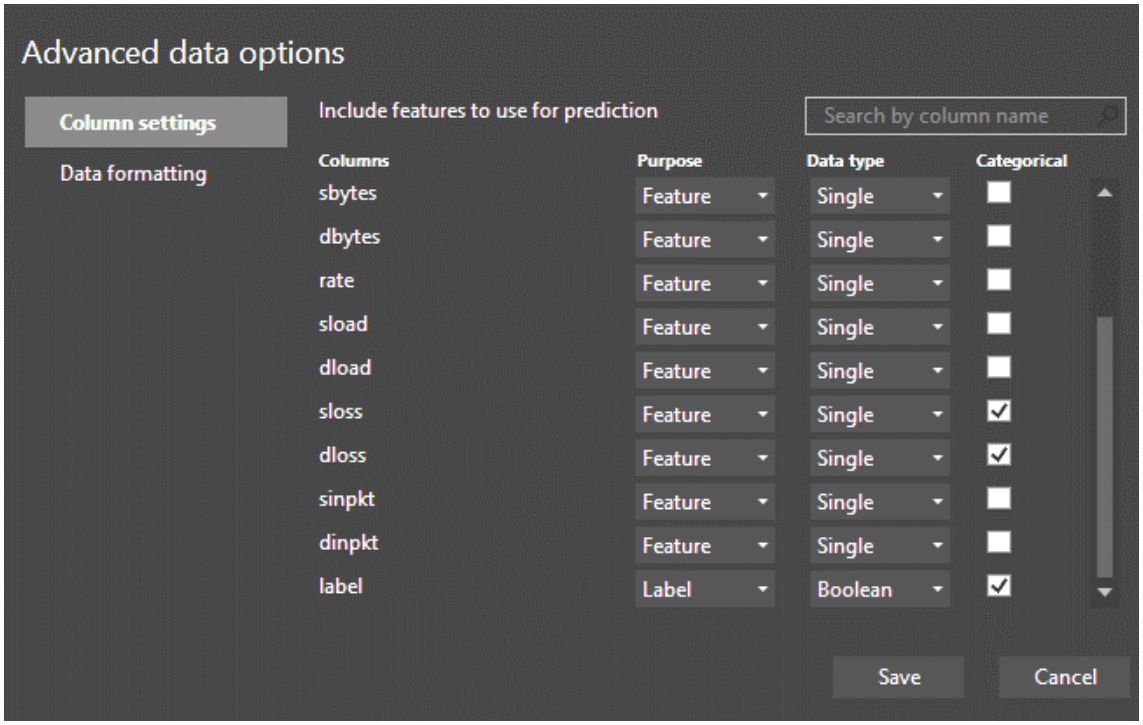


Рисунок 7 – Вкладка «Advanced data options»

После этого можно начинать работу по автоопределению оптимальной модели для наших входных данных. В результате работы «ModelBuilder» получен отчет, изображенный в таблице 4.

Таблица 4 – Отчет «ModelBuilder»

Trainer	Accuracy	AUC	AUPRC	F1-score	Duration	Iteration
FastTreeBinary	0,9290	0,9826	0,9879	0,9351	2,5	45
FastForestBinary	0,9137	0,975	0,982	0,9198	1,1	39
FastTreeBinary	0,8795	0,9188	0,9065	0,9007	3,2	40
FastTreeBinary	0,8643	0,9696	0,9731	0,8891	0,3	36
FastForestBinary	0,8605	0,9613	0,9722	0,883	3,3	20

В данном отчете содержатся поля, описанные ниже.

А. Accuracy – точность. Она рассчитывается как количество правильных прогнозов, деленное на общее количество прогнозов.

Б. AUC – площадь под кривой рабочих характеристик приемника. При прогнозировании класса образца алгоритм машинного обучения сначала вычисляет вероятность того, что обработанный образец принадлежит к определенному классу, и если это значение превышает некоторый предопределенный порог, он помечает этот класс.

Например, для первого образца алгоритм предсказывает, что существует 0,7 (70 %) вероятность того, что это класс 0, а порог равен 0,6, то образец будет помечен как класс 0. Это означает, что для разных порогов можно получить разные метки. Для того чтобы отобразить процент истинных положительных результатов по сравнению с уровнем ложноположительных результатов для различных пороговых значений, была построена ROC-кривая.

Данная метрика используется как мера производительности. Можно сказать, что ROC – это кривая вероятности, а AUC – мера разделимости, т. е. комбинация AUC-ROC означает, что модель способна различать классы. Чем выше это значение, тем лучше.

Пример такой площади изображен на рисунке 8.

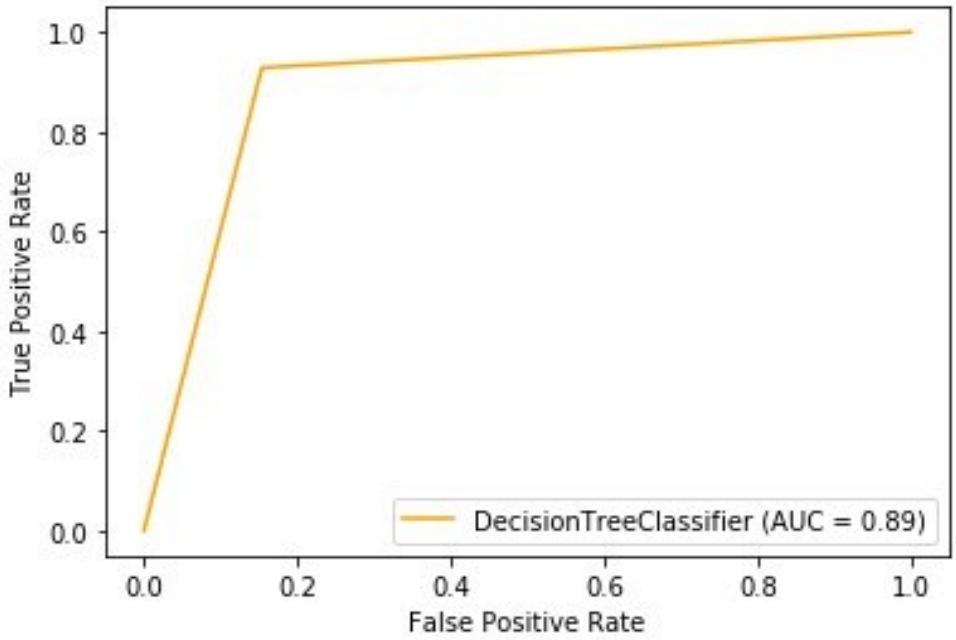


Рисунок 8 – Кривая «AUC-ROC»

В. AUPRC – площадь под кривой точности-полноты. Чтобы правильно оценить модель, необходимо учитывать обе метрики. Кривая точность-полнота показывает компромисс между точностью и полнотой.

Область под кривой представляет как высокую полноту, так и высокую точность. Высокие баллы для обоих показывают, что классификатор возвращает точные результаты с большинством положительных результатов (рис. 9).

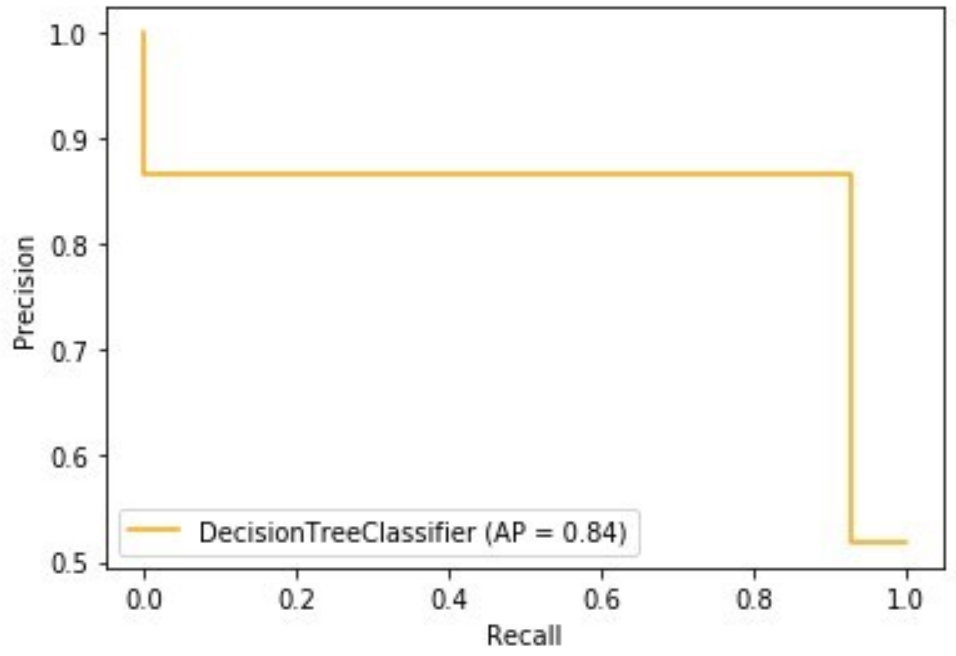


Рисунок 9 – Кривая «AUPRC»

Г. F1 Score- Оценка F1. Является самой популярной метрикой, которая сочетает в себе точность и полноту. Она представляет из себя гармоническое среднее из них. Для бинарной классификации можно определить ее по формуле 1:

$$\text{F1 score} = 2 * \frac{\text{Точность} * \text{Полноту}}{(\text{Точность} + \text{Полнота})}. \quad (1)$$

Первый параметр показывает название алгоритма для обучения, а последние 2 – время на обучение и итерацию соответственно.

В результате данного исследования и проведения сравнения выбранных ранее методов машинного обучения с использованием нейронных сетей можем сделать вывод о целесообразности использования в решении поставленной задачи следующих алгоритмов: «Быстрый лес» и «Быстрое дерево».

Для этого первоначально нужно создать класс, который характеризует пакет для анализа. Строение класса описано в таблице 5.

Таблица 5 – Данные класса, характеризующего данные пакета

Наименование переменной	**Тип**
Dur	float
Proto	string
Service	string
Dpkts	float
Sbytes	Float
Dbytes	float
Rate	float
Sload	float
Dload	float
Sloss	float
Dloss	float
Sinpkt	float
Dinpkt	float
Label	bool

Все параметры были описаны выше, кроме последнего. Он служит для вывода результата работы алгоритма.

Для второй проверки алгоритмов был использован nugget пакет ML.NET.Auto. Данный пакет позволяет произвести те же самые вычисления, что и ModelBuilder, только с более точной настройкой.

В качестве алгоритмов для проверки были выбраны наши фавориты, а в качестве метрики оптимизации выбрана оценка F1. После запуска нашего приложения был получен результат, подтверждающий ожидаемую эффективность выбранных алгоритмов, показанный на рисунке 11.

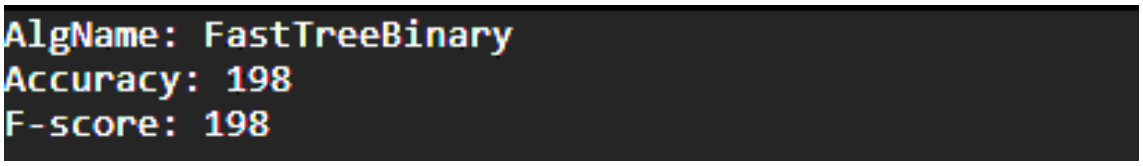


Рисунок 11 – Результат оценки алгоритмов по метрике F1

ЗАКЛЮЧЕНИЕ

1. В исследовании был использован набор данных «UNSW-NB15». Для упрощения и ускорения работы нейросети была поставлена задача бинарной классификации (определение нормального или аномального пакета). Далее получаем небольшой парсер, написанный на языке программирования C#, который конвертирует исходный файл к виду необходимой модели.

2. Были использованы следующие алгоритмы: быстрое дерево, LightGbm, быстрый лес. Далее были проведены 2 теста: грязный, с помощью расширения для ML.NET, под названием ModelBuilder, а также более точный, уже с помощью ручной подгонки.

3. Была построена модель, которая используется как мера производительности. Кривая «AUC-ROC» показывает процент истинных положительных результатов по сравнению с уровнем ложноположительных результатов для различных пороговых значений.

4. Из отчета, приведенного в статье, можно выделить 2 лидера, а именно алгоритмы «Быстрый лес» и «Быстрое дерево».

Список литературы

1. Cynet.com: Network Attacks and Network Security Threats. – URL: https://www.cynet.com /network-attacks/network-attacks-and-network-security-threats/(дата обращения: 27.09.2023).

2. Purplesec.us: Purple Sec’s Cyber Security Maturity Model For Business. – URL: https://purplesec.us/learn/ (дата обращения: 28.09.2023).

3. Mdsny.com: 5 Top Machine Learning Use Cases for Security. – URL: https://www.mdsny.com/5-top-machine-learning-use-cases-for-security/ (дата обращения: 28.09.2023).

4. Stackoverflow.com: Find the best answer to your technical question, help others answer theirs. – URL: https://stackoverflow.com/questions/22500525/nsl-kdd-features-from-raw-live-packets/22522174#22522174 (дата обращения: 27.09.2023).

5. Aarnet: Attack Detection and Mitigation in IoT-Fog Architecture: Handling Class Imbalance Problem. – URL: https://cloudstor.aarnet.edu.au/plus/index.php/s/2DhnLGDdEECo4ys?path=%2FUNSW-NB15%20-%20Reports (дата обращения: 27.09.2023).

6. Платонов, В. В. Обнаружение сетевых атак в компьютерных сетях с использованием методов интеллектуального анализа данных / В. В. Платонов П. О. Семенов // Интеллектуальные технологии на территории. 2016. № 4. – URL: https://cyberleninka.ru/article/n/detecting-network-attacks-in-computer-networks-by-using-data-mining-methods , свободный. – Заглавие с экрана. – Яз. рус. (дата обращения: 27.09.2023).

7. Володин, И. В. Классификация механизмов атак и исследование методов защиты систем с использованием алгоритмов машинного обучения и искусственного интеллекта / И. В. Володин, М. М. Путято, А. С. Макарян, В. Ю. Евглевский // Прикаспийский журнал: управление и высокие технологии. – 2021. – № 2 (54). – С. 91–98.

8. Черкасов, А. Н. Разработка модели распознавания вредоносного программного обеспечения на основе свёрточных нейронных сетей / А. Н. Черкасов, Е. А. Туркин // Инновационное развитие техники и технологий в промышленности (ИНТЕКС-2021) : сборник материалов Всероссийской научной конференции молодых исследователей с международным участием, Москва, 12–15 апреля 2021 года. – Москва : Федеральное государственное бюджетное образовательное учреждение высшего образования «Российский государственный университет имени А.Н. Косыгина (Технологии. Дизайн. Искусство)», 2021. – Т. 4. – С. 72–76.

References

1. *Cynet.com: Network Attacks and Network Security Threats.* Available at: https://www.cynet.com/network-attacks/network-attacks-and-network-security-threats/ (accessed 27.09.2023).

2. *Purplesec.us: Purple Sec's Cyber Security Maturity Model For Business.* Available at: https://purplesec.us/learn/ (accessed 28.09.2023).

3. *Mdsny.com: 5 Top Machine Learning Use Cases for Security.* Available at: https://www.mdsny.com/5-top-machine-learning-use-cases-for-security/ (accessed 28.09.2023).

4. *Stackoverflow.com: Find the best answer to your technical question, help others answer theirs.* Available at: https://stackoverflow.com/questions/22500525/nsl-kdd-features-from-raw-live-packets/22522174#22522174 (accessed 27.09.2023).

5. *Aarnet: Attack Detection and Mitigation in IoT-Fog Architecture: Handling Class Imbalance Problem.* Available at: https://cloudstor.aarnet.edu.au/plus/index.php/s/2DhnLGDdEECo4ys?path=%2FUNSW-NB15%20-%20Reports (accessed 27.09.2023).

6. Platonov, V. V., Semenov, P. O. Detecting network attacks in computer networks by using data Mining methods. *Intelligent technologies in transport*, 2016, no. 4. URL: https://cyberleninka.ru/article/n/detecting-network-attacks-in-computer-networks-by-using-data-mining-methods (accessed 27.09.2023).

7. Volodin, I. V., Putyato, M. M., Makaryan, A. S., Yevglevsky, V. Yu. Classification of attack mechanisms and research of system protection methods using machine learning and artificial intelligence algorithms. Caspian Journal: Control and High Technologies, 2021, no. 2 (54), pp. 91–98.

8. Cherkasov, A. N., Turkin, E. A. Development of a malicious software recognition model based on convolutional neural networks. *Innovative development of technology and technologies in industry (INTEX-2021) : collection of materials of the All-Russian Scientific Conference of Young researchers with international participation, Moscow, April 12–15, 2021.* Moscow, Federal State Budgetary Educational Institution of Higher Education "Kosygin Russian State University (Technologies. Design. Art)", 2021, vol. 4, pp. 72–76.

Статья поступила в редакцию 09.10.2023; одобрена после рецензирования 13.10.2023; принята к публикации 20.10.2023.

The article was submitted 09.10.2023; approved after reviewing 13.10.2023; accepted for publication 20.10.2023.

СИСТЕМНЫЙ АНАЛИЗ, УПРАВЛЕНИЕ И ОБРАБОТКА ИНФОРМАЦИИ

DOI 10.54398/20741707_2023_4_73
УДК 004.89

МЕТОДОЛОГИЯ СБОРА И АНАЛИЗА ДАННЫХ В СОЦИАЛЬНЫХ МЕДИА ТУРКМЕНИСТАНА

Марьенков Александр Николаевич, Астраханский государственный университет имени В. Н. Татищева, 414056, Российская Федерация, г. Астрахань, ул. Татищева, 20а,
кандидат технических наук, доцент, ORCID: 0000-0003-1378-3553, e-mail: marenkovan17@gmail.com
Кривенко Анастасия Ивановна, Астраханский государственный университет имени В. Н. Татищева, 414056, Российская Федерация, г. Астрахань, ул. Татищева, 20а,
магистр, ORCID: 0000-0003-4851-8717, e-mail: krivenkoanastasia33@gmail.com

В статье рассматриваются методологические аспекты сбора и анализа текстовых данных из социальных медиа с целью оценки социальных настроений в Туркменистане по отношению к России. Современные медиа-платформы, включая социальные сети, новостные медиа, блоги и форумы, представляют богатый источник информации о мнениях и взглядах населения, и их анализ может быть полезным для понимания общественных трендов и динамики. Статья также обсуждает технические аспекты сбора данных, обработку текстовых данных и методы анализа, необходимые для успешного исследования социокультурных процессов в Туркменистане. Представленный материал может служить руководством для исследователей и аналитиков, занимающихся анализом данных в контексте социальных медиа и их влияния на общественные настроения.

Ключевые слова: социальные медиа, Туркменистан, сбор данных, анализ текстовых данных, социальные настроения, общественные тренды, общественное мнение, анализ данных

Финансирование: *исследование выполнено за счет гранта Российского научного фонда № 22-18-00301 «Процесс конструирования новых идентичностей в Каспийском макрорегионе в контексте социетальной безопасности».*

METHODOLOGY FOR COLLECTION AND ANALYSIS OF DATA IN SOCIAL MEDIA OF TURKMENISTAN

Marenkov Alexander N., Astrakhan Tatishchev State University, 20a Tatishchev St., Astrakhan, 414056, Russian Federation,
Cand. Sci. (Engineering), ORCID: 0000-0003-1378-3553, e-mail: marenkovan17@gmail.com
Krivenko Anastasiya I., Astrakhan Tatishchev State University, 20a Tatishchev St., Astrakhan, 414056, Russian Federation,
master, ORCID: 0000-0003-4851-8717, e-mail: krivenkoanastasia33@gmail.com

The article discusses the methodological aspects of collecting and analyzing text data from social media in order to assess social sentiment in Turkmenistan. Modern media platforms, including social networks, news media, blogs and forums, provide a rich source of information about the opinions and views of the population, and their analysis can be useful in understanding social trends and dynamics. The article also discusses the technical aspects of data collection, textual data processing, and analysis methods necessary for successful research into sociocultural processes in Turkmenistan. The presented material can serve as a guide for researchers and analysts involved in data analysis in the context of social media and their influence on public sentiment.

Keywords: social media, Turkmenistan, data collection, text data analysis, social sentiment, public trends, public opinion, data analysis

Financial support: *the research was supported by the Russian Science Foundation grant no. 22-18-00301 "The process of constructing new identities in the Caspian macroregion in the context of societal security".*

ВВЕДЕНИЕ

Социальные медиа, СМИ и социальные сети стали неотъемлемой частью современного информационного пространства и оказывают значительное влияние на формирование общественных настроений и восприятие событий. В контексте международных отношений анализ социальных настроений населения становится особенно важным, так как он позволяет понимать, какие факторы и события могут повлиять на взаимодействие между странами. В частности, если рассматривать

территорию Прикаспийского региона, куда входят регионы России, а также Казахстан, Туркменистан, Иран, Азербайджан [1], то можно говорить о попытке формирования «каспийской идентичности» [2] и оценке текущих настроений общества относительно рассматриваемых региональных и глобальных проблем.

Целью данного исследования является создание методологии сбора и анализа данных из социальных медиа, СМИ и социальных сетей Туркменистана с целью выявления социальных настроений населения в отношении России.

Сбор, очистка и анализ данных из социальных медиа, СМИ и социальных сетей Туркменистана позволят выявить динамику и характер социальных настроений населения Туркменистана в отношении России, а также выявить ключевые факторы, оказывающие влияние на эти настроения.

ОБЗОР ЛИТЕРАТУРЫ

Анализ больших данных набирает популярность и значимость, и сейчас большое количество ученых заинтересованы в исследованиях в данной сфере. Анализ больших данных в динамике может раскрывать степень изменения отношений общества к определенным вопросам, оценивать воздействие социальной среды на различные группы людей, особенно в контексте определенных географических регионов, и приближать усилия исследователей по выявлению и разъяснению, как население идентифицирует себя с определенными группами. Исследования проводились в различных направлениях. Отдельно можно выделить подходы по созданию модели машинного обучения для автоматического прогнозирования политических взглядов российских пользователей социальной сети «ВКонтакте» на основе микроподхода к анализу данных [3]. Среди других подходов можно выделить работы, связанные с сентимент-анализом данных социальных сетей [4], исследования в области анализа политических публикаций в интернет-пространстве [5], а также методы сбора и анализа текста в интернете и цифрового следа пользователей [6], которые предлагают полезный инструментарий для работы с текстовыми данными из различных источников. Также были рассмотрены методы сбора больших данных на примере работ российских исследователей о сборе данных в сети Телеграмм [7] и альтернативные методы веб-скрейпинга сайтов для решения возникших проблем в ходе исследования, а именно применение технологии RPA для анализа данных [8].

ОПРЕДЕЛЕНИЕ ИСТОЧНИКОВ ДАННЫХ

Прежде всего необходимо сформулировать цели и задачи исследования. Это поможет определить, какие источники данных наиболее релевантны для достижения целей.

Следует также отметить, что в рамках данного исследования существуют ограничения, связанные с выбором групп социальных сетей и каналов в СМИ. Эти ограничения в значительной степени определяются политическими и социальными контекстами, характерными для Туркменистана.

Сбор данных из социальных медиа, СМИ и социальных сетей является первым этапом исследования. В Туркменистане, несмотря на его ограниченный доступ к мировому интернету и строгий контроль над информацией со стороны правительства, существуют несколько популярных социальных сетей, которые остаются доступными для местных пользователей. Одной из самых популярных социальных сетей в Туркменистане является ВКонтакте. Это социальное средство связи позволяет пользователям общаться, делиться контентом, а также создавать и присоединяться к группам с общими интересами. Еще одной популярной социальной сетью является Facebook. Доступ к Facebook периодически ограничивается правительством, некоторые пользователи все равно находят способы обойти блокировку и оставаться активными на этой платформе. Среди молодежи и студентов популярна сеть Instagram. Что касается иммигрантов из Туркменистана, они часто используют WhatsApp, Telegram и Viber для общения с родственниками и друзьями в Туркменистане. Эти мессенджеры позволяют обмениваться текстовыми сообщениями, аудио- и видеозвонками, что делает их особенно популярными среди мигрантов.

Несмотря на ограничения в доступе к интернету и социальным сетям, пользователи Туркменистана находят способы подключения к мировой сети и продолжают активно общаться и выражать свои мнения в социальных медиа. Это могут быть социальные медиа, СМИ, академические статьи, официальные документы и многие другие (табл.).

Таблица – Новостные Telegram-каналы Туркменистана

Telegram-канал	Количество подписчиков	Ссылка на Telegram-канал
Хроника Туркменистана	4 343	https://t.me/hronikatm
Türkmençilik | Туркменистан	489	https://t.me/turkmenchilik
Turkmen.news Новости Туркменистана	9 955	https://t.me/anthabar
Демократический выбор Туркменистана	121	https://t.me/dwt_kanal
Turkmen Yurt Tv	271	https://t.me/Turkmenyurttv

Исходя из выбранных Telegram-каналов, можно сделать вывод, что их аудитория обычно отличается небольшим размером и редко принадлежит оппозиционным группам или организациям. Эти каналы, вероятно, ориентированы на узкую аудиторию, имеющую конкретные интересы или точку зрения. Важно отметить, что степень оппозиционности или политической ориентации каналов может различаться, и анализ их содержания и влияния на аудиторию может быть ключевой частью исследования в контексте социокультурных процессов и общественных настроений в Туркменистане.

ВЫБОР КЛЮЧЕВЫХ СЛОВ

Следующим этапом сбора данных является выбор ключевых слов и фраз, связанных с отношениями между Туркменистаном и Россией. Эти ключевые слова используются для мониторинга социальных медиа, новостных сайтов и социальных сетей. Также используются геолокационные теги, чтобы отслеживать контент, созданный в Туркменистане.

Примером ключевых слов, используемых для сбора данных, может служить составленный запрос:

«Отношения (Туркмении, Туркменистана), (Туркменистан, Туркмения, туркмены)* (русский язык, русские классы, русские), (памятники, монументы)* (Туркменской ССР, советской эпохи, СССР, советского союза), теке, кизил-арватский, марыйский, балканский, чарджоуский, ташаузский, ахалский»*

Данный запрос был использован в программе Brand Analytics [9] для целей анализа отношений Туркмении к России. Этот комплексный запрос включает в себя разнообразные ключевые слова и фразы, охватывая различные аспекты и периоды времени, что позволяет более полно исследовать и анализировать социальные настроения и отношения между этими двумя странами. В поисковом запросе учитываются исторические и культурные аспекты (например, «советская эпоха», «монументы»), что позволяет углубиться в контекст и прошлое отношений между странами. Упоминание регионов (например, «Теке», «Марыйский») позволяет учитывать различия в мнениях и отношениях на разных территориях Туркменистана.

Используя данный запрос, можно анализировать тональность высказываний, содержащих указанные ключевые слова, чтобы определить, какие настроения и отношения преобладают в обсуждениях.

СБОР ДАННЫХ

Существует несколько методов сбора данных из социальных сетей и СМИ. Так, например, многие социальные сети предоставляют API, которые позволяют разработчикам получать доступ к данным платформы. Это наиболее надежный и официальный способ сбора данных. Например, Telegram API, Facebook Graph API, Instagram Graph API и др. Однако для доступа к API могут потребоваться разрешения и ключи, и могут существовать ограничения по объему данных. На рисунке 1 представлен фрагмент кода соединения с аккаунтом разработчика Telegram API для сбора данных.

Сбор данных из Telegram-каналов при помощи Telegram API осуществляется по открытой структуре каналов с помощью автоматизированных средств сбора информации, что делает этот источник данных доступным для анализа.

```
import configparser
import json
import os
import asyncio
from datetime import date, datetime

from telethon import TelegramClient
from telethon.errors import SessionPasswordNeededError
from telethon.errors import MsgIdInvalidError
from telethon.tl.functions.messages import (GetHistoryRequest)
from telethon.tl.types import (
    PeerChannel
)

class DateTimeEncoder(json.JSONEncoder):
    def default(self, o):
        if isinstance(o, datetime):
            return o.isoformat()

        if isinstance(o, bytes):
            return list(o)

        return json.JSONEncoder.default(self, o)


# Reading Configs
config = configparser.ConfigParser()
config.read("config.ini")

# Setting configuration values
api_id = config['Telegram']['api_id']
api_hash = config['Telegram']['api_hash']
api_hash = str(api_hash)

phone = config['Telegram']['phone']
username = config['Telegram']['username']
```

Рисунок 1 – Пример соединения с аккаунтом разработчика Telegram API

Также практичным методом сбора информации является веб-скрейпинг [10]. Этот метод включает в себя автоматизированный сбор данных с веб-страниц социальных сетей. Необходимо отметить, что в некоторых случаях разработчики веб-сайтов принимают меры для защиты от веб-скрейпинга и парсинга данных. Эти меры могут включать в себя различные технические и юридические методы, предназначенные для ограничения или предотвращения автоматизированного сбора информации. В таких случаях использование программных роботов (RPA) может стать более эффективным способом сбора данных. RPA использует программные роботы для выполнения рутинных задач, таких как сбор и анализ данных. RPA-роботы могут быть настроены на сбор данных из Facebook, включая посты, комментарии, лайки и другую информацию. На рисунке 2 представлен пример кода сбора данных из социальной сети Facebook с применением RPA.

```python
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
import pyautogui
from bs4 import BeautifulSoup
import time
from lxml import html
import string
from googletrans import Translator


driver = webdriver.Chrome()
driver.get("https://m.facebook.com/login/") #  https://m.facebook.com/login/
email = "" #логин для входа
password = "" #пароль для входа

email_xpath = '//*[@id="m_login_email"]'
pass_xpath = '//*[@id="m_login_password"]'
but_xpath = '//*[@id="login_password_step_element"]/button'

email_element = driver.find_element_by_xpath(email_xpath)
pass_element = driver.find_element_by_xpath(pass_xpath)
but_element = driver.find_element_by_xpath(but_xpath)

email_element.send_keys(email)
pass_element.send_keys(password)
but_element.click()

time.sleep(4)
```

Рисунок 2 – Отрывок RPA сбора данных для Facebook

Однако при использовании RPA для сбора данных с веб-сайтов, защищенных от сбора, необходимо соблюдать законы и политики использования данных.

В ходе исследования был осуществлен сбор данных из различных источников, включая социальные медиа, СМИ и социальные сети Туркменистана, в период с 02.01.2022 по 31.07.2023. Общий объем собранных данных составил 747771 строк, обладающих следующей структурированной характеристикой: дата, ID сообщения, заголовок, текст, источник, URL, автор, место публикации, комментарии, репосты, лайки, просмотры, дубли, тональность, страна, регион, город, язык.

После процесса тщательной очистки и обработки данных было сохранено 272033 качественных записей. Основными источниками собранных данных оказались социальные сети и мессенджеры, включая Вконтакте, Одноклассники, Telegram, Facebook и Instagram, а также разнообразные новостные сайты и интернет-СМИ.

ОЧИСТКА ДАННЫХ

Собранные данные подвергаются процессу очистки, включая удаление дубликатов, исправление опечаток и фильтрацию символов. Этот этап важен для получения точных и надежных данных. Данные могут быть очищены как вручную, так и при помощи машинных методов.

Машинная очистка данных представляет собой процесс автоматизированной обработки и преобразования собранных текстовых данных с целью улучшения их качества и структуры. Машинные алгоритмы могут идентифицировать и удалять повторяющиеся записи или сообщения, что помогает избежать излишней дубликации информации. Так, например, автоматическая коррекция опечаток и грамматических ошибок может быть применена для улучшения читаемости текста и точности анализа. Машинные алгоритмы могут удалить лишние символы, специфичные знаки и символы форматирования, такие как хэштеги, URL-адреса или эмодзи. Текст может быть разбит на токены (слова или фразы) и подвергнут лемматизации (приведение слов к базовой форме), чтобы унифицировать данные и улучшить точность анализа.

Использование регулярных выражений, и Perl Data Language (PDL) запросов также предоставляет мощный инструмент для поиска и модификации текстовых данных. В контексте исследования

регулярные выражения могут быть применены для поиска и извлечения текста, соответствующего определенным шаблонам, что полезно для выделения ключевой информации. Регулярные выражения могут быть использованы для замены или фильтрации определенных фрагментов текста, что помогает улучшить чистоту данных.

Очистка данных с использованием Structured Query Language for Textual Data (SRL) запросов представляет собой эффективный способ выделения и фильтрации текстовых записей из набора данных, исходя из заданных ключевых слов или фраз. Ниже приведен пример SRL-запроса, применяемый в исследовании:

Singleroot (трилингвизм) @[Текст] or singleroot (языковая политика) @[Текст] or singleroot (русский язык) @[Текст] or singleroot (казахский язык) @[Текст] or singleroot (английский язык) @[Текст] or singleroot (языковые патрули) @[Текст] or singleroot (латиница) @[Текст] or singleroot (кириллица) @[Текст] or singleroot (языковая реформа) @[Текст]

Приведенный SRL-запрос позволяет выявить записи, соответствующие ключевым словам и фразам, связанным с лингвистической и языковой политикой.

Применение машинной очистки данных, методов очистки PDL и SRL позволяет обработать и предварительно подготовить текстовые данные для дальнейшего анализа. Этот этап является важным шагом в обеспечении качества и точности анализа отношений Туркменистана к России на основе данных из социальных сетей и СМИ.

АНАЛИЗ ДАННЫХ

Основным этапом данного исследования является анализ данных. Для более глубокого понимания социокультурных процессов и формирования объективных выводов используются методы анализа текстовых данных, основанных на естественном языке.

В рамках данной статьи применялось программное обеспечение PolyAnalyst [11]. Применение данного программного обеспечения позволило получить более глубокое и информативное представление о социокультурных процессах и отношениях между Туркменистаном и Россией. Пример проекта PolyAnalyst представлен на рисунке 3.

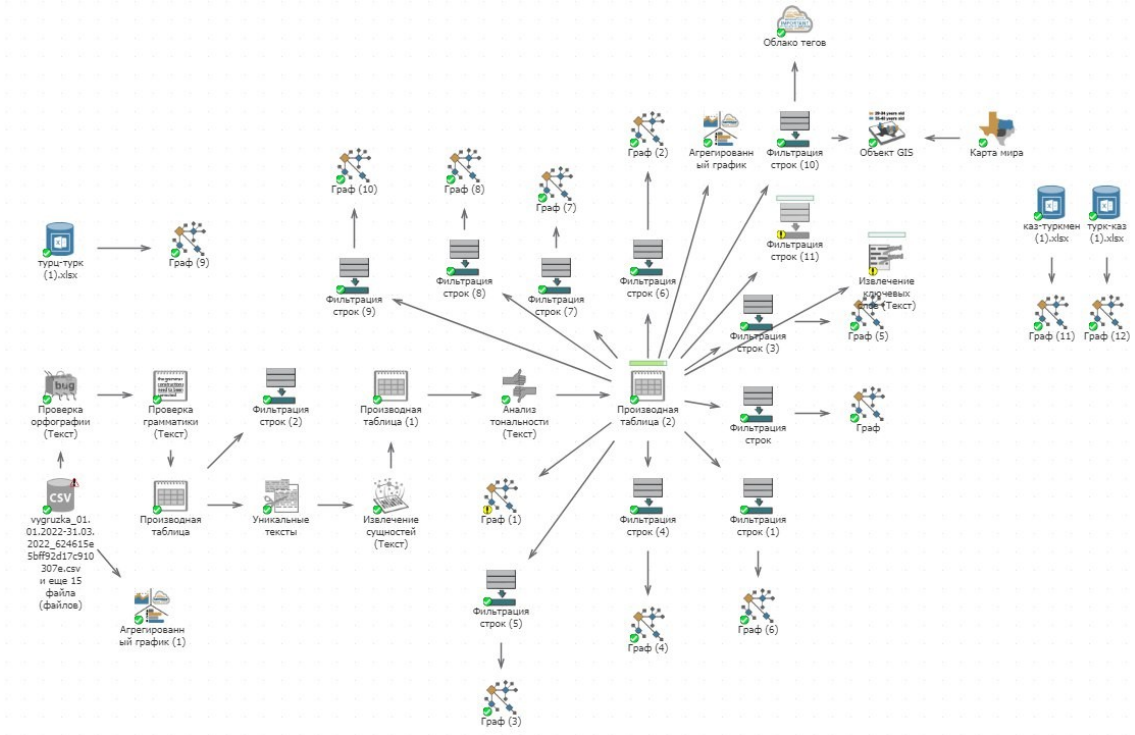


Рисунок 3 – Проект в программе PolyAnalyst

В начале анализа текстовые данные подвергаются структурированию и преобразованию в более удобный для обработки формат, который включает в себя токены (слова и фразы). Далее применяются разнообразные методы, направленные на извлечение ценной информации и понимание настроений.

Один из ключевых этапов в анализе данных – это выделение тематических областей и основных топиков, обсуждаемых в текстах. Такие темы могут включать в себя политические события, экономические аспекты, культурные обмены и, конечно же, отношения между Туркменистаном и Россией. Идентификация и классификация тем позволяет более глубоко исследовать содержание текстовых данных и выявить ключевые направления обсуждений.

Определение тональности текстовых данных – важный аспект исследования. Сентимент-анализ позволяет определить, являются ли высказывания в тексте положительными или негативными. Это необходимо для оценки общих настроений населения в отношении России и понимания, какие события или явления влияют на эти настроения.

Сетевой анализ полезен для поиска различных участников и сообществ, влияющих на социальные настроения [12]. Он включает в себя определение ключевых пользователей, групп и сообществ, активно обсуждающих отношения между Туркменистаном и Россией. Анализ взаимосвязей и влияния между участниками и сообществами может помочь выявить факторы, формирующие социальные настроения.

Анализ данных во времени позволяет выявить динамику социальных настроений. Изучение изменений в обсуждении тем со временем помогает выявить события или факторы, которые могут влиять на отношения между Туркменистаном и Россией. Этот анализ может также выявить сезонные и событийные тренды, способствующие пониманию долгосрочной динамики.

Анализ географических данных позволяет исследовать, какие регионы активнее всего обсуждают отношения между Туркменистаном и Россией. Это включает в себя выделение географических ключевых слов и местоположений, связанных с этими двумя странами.

ВИЗУАЛИЗАЦИЯ ДАННЫХ

Визуализация данных выполнялась с помощью PolyAnalyst. Программа предоставляет возможность наглядно представить результаты анализа. Были созданы различные типы графиков и диаграмм для визуализации результатов. На рисунке 4 приведены примеры полученной визуализации данных.

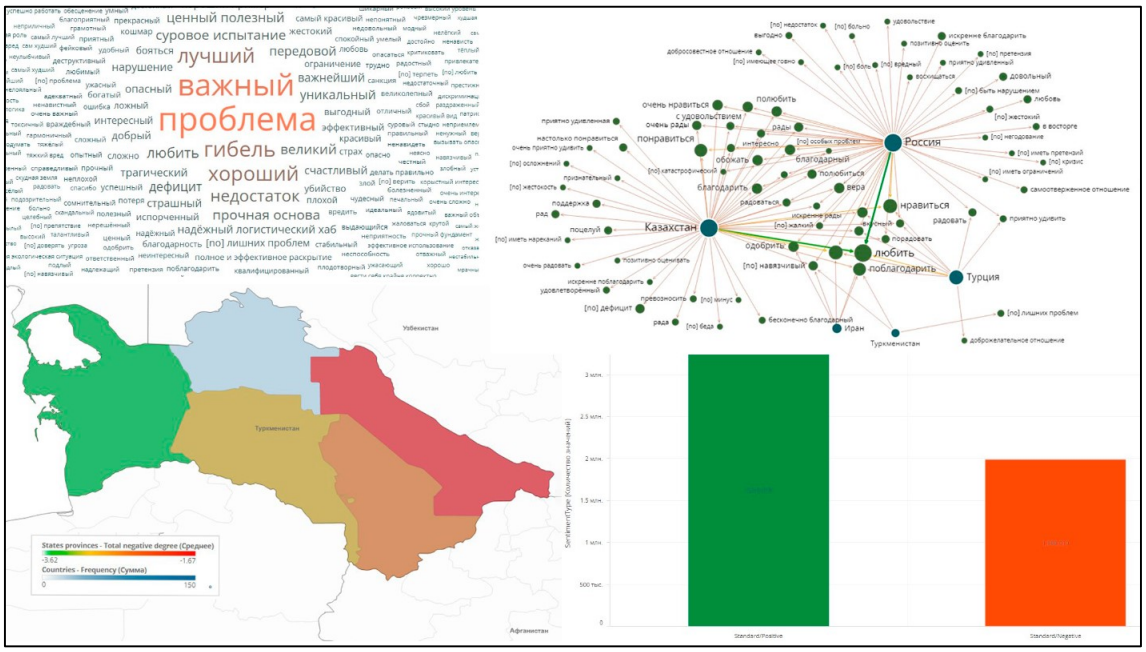


Рисунок 4 – Визуализация данных в программе PolyAnalyst

Облака тегов используются для наглядного представления наиболее часто встречающихся слов или фраз в текстовых данных. Для данных, которые содержат географическую информацию, позволяют создать карты с отметками и цветовыми кодировками для отображения распределения событий или настроений по регионам. Полезной формой визуализации является таблица или список, представляющие ключевые слова, темы или пользователей с соответствующими метриками, такими как частота упоминания или тональность.

Если анализируются взаимосвязи между участниками и сообществами, PolyAnalyst может создавать сетевые диаграммы для наглядного представления связей и влияния между ними.

Для анализа динамики данных с течением времени строятся временные графики, отражающие изменения сентимента или частоту обсуждений.

Итоговые результаты могут быть вынесены на интерактивные дашборды, которые объединяют несколько видов визуализации в одном месте, что позволяет исследователям более удобно исследовать и интерпретировать данные.

Визуализация данных с помощью PolyAnalyst помогает лучше понять результаты анализа и выявлять важные паттерны и тенденции в текстовых данных. Эти визуализации могут быть включены в отчеты и презентации для наглядного представления результатов исследования.

ВЫВОДЫ

Рассмотрим выводы, которые можно сделать по каждой из вышеупомянутых визуализаций данных, изображенных на рисунке 4. Тепловая карта, отображающая изменение позитивных и негативных настроений в разных районах страны, может помочь выявить географические различия в общественных мнениях. Например, красные зоны (негативные настроения) могут свидетельствовать о наличии серьезных проблем или конфликтов в этих регионах. Зеленые зоны (позитивные настроения) могут указывать на стабильность и удовлетворенность.

Частота слов, связанных с «проблемами» и «недостатками», может указывать на важные социальные или экономические проблемы, которые требуют внимания и решения. Появление слов, таких как «поблагодарить», «хороший» и «надежный», может указывать на успешные и положительные аспекты общественной жизни.

Из графа связей между странами Прикаспийского региона можно сделать вывод, что наиболее часто упоминаются положительные слова, такие как «одобрить», «любить» и «поблагодарить». Это может указывать на укрепление дружественных отношений между Туркменистаном и соседними странами, что может иметь важное значение в контексте международной дипломатии и торговых отношений.

Кроме вышеупомянутых методов, также возможно провести анализ лидеров мнений, что позволяет выявить ключевых персон, чьи мнения и сообщения оказывают наибольшее влияние на общественные настроения и дискуссии. Также возможно отслеживание динамики настроений по определенным тематикам в ретроспективе, что позволяет провести долгосрочный анализ изменений в общественных настроениях. Выявление эволюции восприятия конкретных тем и событий может помочь исследователям определить факторы, влияющие на эти изменения. Выделение ключевых тематик текстов дает возможность лучше понимать, о чем именно говорится в обсуждениях. Это помогает в определении приоритетов и акцентов в сообщениях и обсуждениях, а также позволяет исследователям более точно определить темы, которые оказывают влияние на общественные настроения.

Визуализация данных и анализ социокультурных процессов в СМИ могут помочь аналитикам мониторить изменения в политических настроениях и культурной идентичности. На основе этих данных можно разрабатывать стратегии и политику, учитывая общественные предпочтения и потребности.

ЗАКЛЮЧЕНИЕ

Исследование, проведенное в статье, подчеркивает значимость анализа данных из социальных медиа и СМИ для понимания динамики международных отношений и социокультурных процессов. Разработанная методология может быть применена и в других регионах для анализа отношений между странами, что делает ее ценным инструментом для исследования и прогнозирования социокультурных и политических явлений.

Для успешной реализации данной методологии необходимо обладать высокой вычислительной мощностью и доступом к большой базе данных, содержащей текстовую информацию из социальных медиа и СМИ. Эффективные алгоритмы NLP также играют ключевую роль в обработке больших объемов данных.

Анализ данных социальных медиа и СМИ может быть полезен для мониторинга и анализа международных отношений и политических событий. Он позволяет выявлять общественное мнение и реакцию на внешнеполитические события. Государственные и негосударственные организации могут использовать такие исследования для мониторинга общественного мнения по важным социальным и политическим вопросам. Ученые и исследователи могут использовать такие исследования для изучения социокультурных явлений и процессов.

Библиографический список

1. Аманов, М. Э. Основные экологические проблемы Каспийского региона / М. Э. Аманов, К. К. Акмурадова // Каспий и глобальные вызовы. – 2022. – С. 9–14.

2. Baeva, L.V. Challenges to frontier allegories: The Caspian Sea region in Southern Russia / L. V., Baeva, A. P. Romanova // Cultura. International Journal of Philosophy of Culture and Axiology. – 2015. – Vol. 12, № 1. – P. 159–172.

3. Kozitsin, I. V. Modeling Political Preferences of Russian Users Exemplified by the Social Network Vkontakte / I. V. Kozitsin, A. G. Chkhartishvili, A. M. Marchenko, D. O. Norkin, I. A. Osipov et al. // Mathematical Models and Computer Simulations. – 2020. – Vol. 12, № 2. – P. 185–194.

4. Богданов, А. Л. Сентимент-анализ коротких русскоязычных текстов в социальных медиа / А. Л. Богданов, И. С. Дуля // Вестник Томского государственного университета. – Экономика. – 2019. – № 47. – С. 220–241.

5. Boldyreva, A. Internet Queries as a Tool for Analysis of Regional Police Work and Forecast of Crimes in Regions / A. Boldyreva, M. Alexandrov, O. Koshulko, O. Sobolevskiy // Advances in Computational Intelligence. – 2016. – Vol. 1, № 15. – P. 290–301.

6. Харковчук, А. Э. Составление цифрового профиля человека на основе поиска информации по его фотографии из открытых источников в сети интернет / А. Э. Харковчук, Д. Ж. Корзун // Цифровые технологии в образовании, науке, обществе : материалы XIII Всероссийской научно-практической конференции, Петрозаводск, 17–20 сентября 2019 года. – Петрозаводск : Петрозаводский государственный университет, 2019. – С. 199–202.

7. Карабак, И. И. Парсинг телеграм-каналов как элемент системы автоматизированного анализа информации, полученной из сети Интернет / И. И. Карабак, К. А. Зорин, И. М. Ажмухамедов // Прикаспийский журнал: управление и высокие технологии. – 2022. – № 1 (57). – С. 9–17.

8. Plattfaut, R. The Critical Success Factors for Robotic Process Automation / R. Plattfaut, V. Borghoff, M. Godefroid et al. // Computers in Industry. – 2022. – Vol. 138. – P. 103646. – DOI: 10.1016/j.compind.2022.103646.

9. Brand Analytics. – URL: https://brandanalytics.ru/ (дата обращения: 09.10.2023).

10. Diouf, R. Web scraping: state-of-the-art and areas of application / R. Diouf // 2019 IEEE International Conference on Big Data (Big Data). – IEEE, 2019. – С. 6040–6042.

11. Data and Text Analysis Software & Solutions // Megaputer. – URL: https://www.megaputer.com/ (дата обращения: 10.10.2023).

12. Nandwani, P. A review on sentiment analysis and emotion detection from text / P. Nandwani, R. Verma // Social Network Analysis and Mining. – 2021. – Т. 11, № 1. – С. 81.

References

1. Amanov, M. E., Akmuradova, K. K. The main environmental problems of the Caspian region. *The Caspian Sea and Global Challenges*, 2022.

2. Baeva, L. V., Romanova, A. P. Challenges to Frontier Allegories: the Caspian Sea Region in Southern Russia. *Cultura*, 2015, vol. 12, no. 1, pp. 159–172.

3. Kozitsin, I. V., Chkhartishvili, A. G., Marchenko, A. M., Norkin, D. O., Osipov et al. Modeling political preferences of Russian users exemplified by the social network Vkontakte. *Mathematical Models and Computer Simulations*, 2020, vol. 12, no. 2, pp. 185–194.

4. Bogdanov, A. L., Dulya, I. S. Sentiment analysis of short Russian-language texts in social media. *Bulletin of Tomsk State University. Economics*, 2019, no. 47, pp. 220–241.

5. Boldyreva, A., Alexandrov, M., Koshulko, O., Sobolevskiy, O. Internet Queries as a Tool for Analysis of Regional Police Work and Forecast of Crimes in Regions. *Advances in Computational Intelligence*, 2016, vol. 1, no. 15, pp. 290–301.

6. Kharkovchuk, A. E., Korzun, D. Z. Compiling a digital profile of a person based on searching for information from his photograph from open sources on the Internet. *Digital technologies in education, science, society ; proceedings of XIII All-Russian scientific practical conference, Petrozavodsk, September 17–20, 2019.* Petrozavodsk, Petrozavodsk State University, 2019, pp. 199–202.

7. Karabak I. I., Zorin K. A., Azhmukhamedov I. M. Parsing telegram channels as an element of a system for automated analysis of information received from the Internet. *Caspian Journal: Management and High Technologies*, 2022, no. 1 (57), pp. 9–17.

8. Plattfaut, R., Borghoff, V., Godefroid, M. et al. The Critical Success Factors for Robotic Process Automation. *Computers in Industry*, 2022, vol. 138, p. 103646. DOI: 10.1016/j.compind.2022.103646.

9. Brand Analytics. Available at: https://brandanalytics.ru/ (accessed 09.10.2023).

10. Diouf, R. Web scraping: state-of-the-art and areas of application. *2019 IEEE International Conference on Big Data (Big Data).* IEEE, 2019, pp. 6040–6042.

11. Data and Text Analysis Software & Solutions. *Megaputer*. Available at: https://www.megaputer.com/ (accessed 10.10.2023).

12. Nandwani, P., Verma, R. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 2021, vol. 11, no. 1, p. 81.

Статья поступила в редакцию 16.10.2023; одобрена после рецензирования 16.10.2023; принята к публикации 19.10.2023.

The article was submitted 16.10.2023; approved after reviewing 19.10.2023; accepted for publication 19.10.2023.

УДК 004.054

ИССЛЕДОВАНИЕ БАЗОВЫХ АТАК И КОМПРОМЕНТАЦИИ ДОМЕННОЙ WINDOWS-ИНФРАСТРУКТУРЫ

Котов Иван Юрьевич, МИРЭА – Российский технологический университет, 119454, Российская Федерация, г. Москва, пр. Вернадского, 78,

студент, ORCID: 0009-0001-9611-3952, e-mail: vanos03@mail.ru

Брысин Андрей Николаевич, МИРЭА – Российский технологический университет, 119454, Российская Федерация, г. Москва, пр. Вернадского, 78; Институт машиноведения им. А.А. Благонравова Российской академии наук, 101000, Российская Федерация, г. Москва, Малый Харитоньевский переулок, 4,

кандидат технических наук, доцент, ORCID: 0000-0001-8870-5037, e-mail: brysin@mirea.ru

Журавлева Юлия Алексеевна, МИРЭА – Российский технологический университет, 119454, Российская Федерация, г. Москва, пр. Вернадского, 78; Национальный исследовательский университет «МЭИ», 111250, Российская Федерация, г. Москва, Красноказарменная улица, 14, стр. 1.

кандидат технических наук, доцент, ORCID: 0000-0003-3919-5127, e-mail: ulypil@mail.ru

Обеспечение информационной безопасности является важной задачей в наше время. Актуальные угрозы включают в себя разнообразные атаки, направленные на сетевые системы и данные. Одной из ключевых составляющих безопасности Windows-инфраструктуры является понимание актуальных уязвимостей и их эксплуатация. Системы, работающие на операционной системе Windows, подвержены различным угрозам, и администраторам необходимо постоянно обновлять их и внимательно следить за уязвимостями. В статье описаны основные характеристики системы CVE, приведен анализ уязвимостей для доменной Windows-инфраструктуры. В ходе работы была разработана модель тестового стенда, с помощью которой было осуществлено тестирование базовых атак и проведена эксплуатация уязвимостей контроллера домен. Модель включает Kali Linux, Windows 10 pro, Windows Server 2016 (англоязычная версия), Server 2016 (русскоязычная версия).

Ключевые слова: информационная безопасность, серверная система, протокол, шифрование, тестирование

RESEARCH OF BASIC ATTACKS AND COMPROMTION OF WINDOWS DOMAIN INFRASTRUCTURE

Kotov Ivan Y., MIREA – Russian Technological University, 78 Vernadsky Ave., Moscow, 119454, Russian Federation,

student, ORCID: 0009-0001-9611-3952, e-mail: vanos03@mail.ru

Brysin Andrey N., MIREA – Russian Technological University, 78 Vernadsky Ave., Moscow, 119454, Russian Federation; Mechanical Engineering Research Institute of the Russian Academy of Sciences, 4 Maly Kharitonyevsky Pereulok, Moscow, 101000, Russian Federation,

Cand. Sci. (Engineering), Associate Professor, ORCID: 0000-0001-8870-5037, e-mail: brysin@mirea.ru

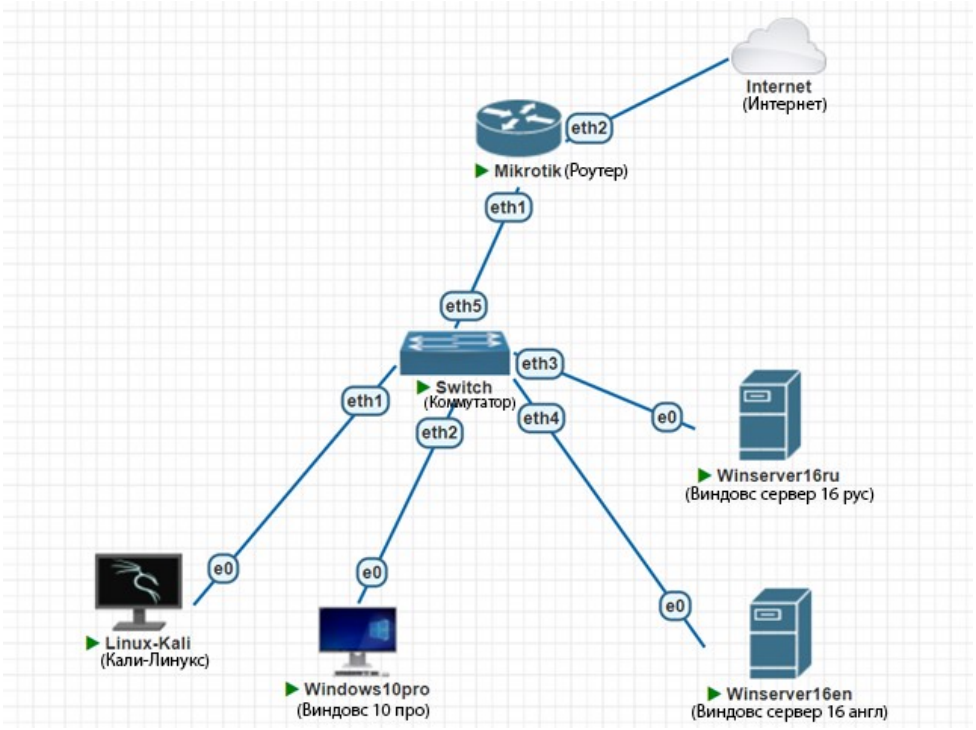
Zhuravleva Yulia A., MIREA – Russian Technological University, 78 Vernadsky Ave., Moscow, 119454, Russian Federation,

Cand. Sci. (Engineering), Associate Professor, ORCID: 0000-0003-3919-5127, e-mail: ulypil@mail.ru

Ensuring information security is an important task in our time. Current threats include a variety of attacks aimed at network systems and data. One of the key components of Windows infrastructure security is understand current vulnerabilities and their exploitation. Systems running on the Windows operating system are susceptible to various threats, and administrators need to constantly update them and keep a close eye on vulnerabilities. The article describes the main characteristics of the CVE system and provides an analysis of vulnerabilities for the Windows domain infrastructure. During the work, a test bench model was develop, with the help of which basic attacks were tested and domain controller vulnerabilities were exploited. The model includes Kali Linux, Windows 10 pro, Windows Server 2016 (English version), Server 2016 (Russian version).

Keywords: information security, server system, protocol, encryption, testing

Graphical annotation (Графическая аннотация)



ВВЕДЕНИЕ

В настоящее время используется много ресурсов на усиление защиты серверных систем от предполагаемых угроз. Одним из наиболее распространенных решений для серверов является доменная Windows-инфраструктура. Доменная Windows-инфраструктура обеспечивает централизованное управление ресурсами компьютера, включая процессор, память, диски, а также обеспечивает доступ к различным приложениям и сервисам через единую точку входа. После волны кибератак на многие организации возникла неотложная потребность в изучении основных методов атак и уязвимостей, которые могут быть использованы злоумышленниками для несанкционированного доступа к системе или ее данным. В доменной Windows-инфраструктуре существует несколько видов угроз безопасности, связанных с использованием данной инфраструктуры: почтовый фишинг (когда злоумышленники пытаются завладеть конфиденциальной информации путем отправки поддельных электронных писем), вирусное ПО на съемных носителях информации (вредоносное программное обеспечение, которое распространяется через съемные устройства, такие как USB-накопители или флеш-карты. Оно может быть установлено на компьютер без согласия пользователя и способно нанести значительный ущерб системе), проникновение на объект (несанкционированный доступ к защищенным объектам, таким как предприятия или учреждения, может произойти из-за недостаточной охраны или из-за ошибок в системах безопасности), уязвимости систем (слабые места в программном обеспечении или операционной системе, которые могут быть использованы злоумышленниками для несанкционированного доступа к системе).

АНАЛИЗ УЯЗВИМОСТЕЙ ДЛЯ WINDOWS

К основным характеристикам системы CVE относятся:

– уникальность идентификаторов: каждая зарегистрированная уязвимость в системе имеет уникальный идентификатор, который состоит из цифр и букв;

– беспристрастность и независимость: CVE не связана с какой-либо конкретной организацией или производителем, является нейтральной системой, служащей интересам всего сообщества информационной безопасности;

– централизованная база данных: CVE поддерживает централизованную базу данных, где собирается, хранится и публикуется информация об уязвимостях. Эта база данных доступна для разработчиков, системных администраторов и исследователей для отслеживания и устранения уязвимостей;

– согласованная нумерация: идентификаторы CVE назначаются и согласовываются организацией, которая работает с сообществом безопасности и производителями для обеспечения правильной нумерации уязвимостей;

– использование в системах оценки рисков: идентификаторы CVE широко используются в инструментах оценки рисков информационной безопасности, таких как системы управления уязвимостями;

– сотрудничество и обмен информацией: CVE способствует обмену информацией о безопасности между организациями и сообществом информационной безопасности, что способствует более эффективной реакции на уязвимости и снижению рисков;

– комплементарность другим стандартам и системам: CVE не заменяет другие системы идентификации уязвимостей, такие как Common Vulnerability Scoring System (CVSS) или National Vulnerability Database (NVD), а дополняет их, предоставляя уникальные идентификаторы уязвимостей.

К ключевым параметрам уязвимости относят: номер уязвимости, шкалу критичности (метрика), вектор, конкретный недостаток реализации, применимость (CPE) и наличие исправления или способы смягчения последствий. Шкала критичности предназначена для того, чтобы быстро понять, насколько критична уязвимость. Метрики (CVSS) служат для определения критичности.

Правила формирования векторов атаки (вектор атаки (AV), сложность атаки (AC), уровень привилегий (PR), взаимодействие с пользователем (UI), влияние на другие компоненты системы (S), влияние на конфиденциальность (C), влияние на целостность (I), влияние на доступность (A)) подробно рассмотрены в [1].

Стоит заметить, что CVE является фундаментальным инструментом для обеспечения безопасности информационных систем. Она позволяет эффективно управлять рисками и реагировать на уязвимости, обеспечивая согласованность в сообществе информационной безопасности.

К актуальным уязвимостям для инфраструктуры Windows относят:

– Zerologon представляет собой ошибку в реализации протокола шифрования AES-CFB8 в Netlogon Remote Protocol (MS-NRPC). Позволяет неаутентифицированному атакующему удаленно получить права администратора домена;

– SMBGhost позволяет неавторизированному атакующему удаленно выполнить код на целевом сервере;

– Proxylogon позволяет внешнему злоумышленнику обойти механизм аутентификации в Microsoft Exchange;

– PrintNightmare представляет собой уязвимость в диспетчере печати;

– Path-the-hash позволяет атакующему получить доступ к удаленному серверу, который использует протоколы аутентификации NTLM (New Technology LAN Manager) или LM (LAN Manager);

– LLMNR (протокол) позволяет компьютерам выполнять разрешение имен хостов в локальной сети;

– NetBIOS (протокол) позволяет обнаружить компьютеры в сети, построенной на базе TCP/IP [2].

МОДЕЛЬ ТЕСТОВОГО СТЕНДА

В ходе работы была разработана и опробована модель тестового стенда (рис. 1), которая включает составляющие: Kali Linux, Windows 10 pro (далее PC1), Windows Server 2016 (англоязычная версия) (далее DC1), Server 2016 (русскоязычная версия) (далее DC2).

Первым шагом было создание виртуальных машин, которые представляли собой различные устройства и компоненты сети. В настоящее время существует широкий спектр образов ОС, таких как Cisco Windows, Linux и др., для создания виртуальных маршрутизаторов и коммутаторов, что позволило построить сетевую топологию, которая соответствовала тестовым потребностям. Далее были подключены виртуальные машины к друг другу с помощью Pnetlab, осуществлена настройка интерфейсов и адреса, проведены различные сетевые тесты и упражнения. Один из важных аспектов Pnetlab позволяет создавать сценарии для тестирования и автоматизации операций сети, что делает процесс более эффективным и позволяет экономить время. Pnetlab также предоставляет инструменты для мониторинга и анализа сетевого трафика, что может быть полезно при отладке и изучении различных аспектов сетевых технологий.

ТЕСТИРОВАНИЕ БАЗОВЫХ АТАК

Все данные каталога Active Directory хранятся в БД в файле ntds.dit (далее NTDS). NTDS используется для аутентификации и авторизации пользователей, а также для управления доступом к ресурсам и разрешения конфликтов при изменении данных [3, 4].

Анализ NTDS проведен с целью исследования структуры базы аутентификационых данных для объектов домена. Когда пользователь создает учетную запись в NTDS Windows, ему обычно предлагается ввести пароль для этой учетной записи. Пароль хранится в файле конфигурации ntlds.config в зашифрованном виде вместе с настройками безопасности, такими как логин и пароль пользователя. В ходе работы была создана копия NTDS для последующей его передаче на линукс для анализа, папка с NTDS была переслана на линукс через протокол smb.

Анализ базы NTDS был осуществлен с помощью утилиты impacket (impacket – это набор программ, написанных на языке Python для работы с сетевыми протоколами, и в первую очередь с Active Directory, установлен в Kali). После запуска анализа были сформированы наборы случайных символов, которые генерируются на основе текстового пароля с помощью специального алгоритма (хэширования) всех пользователей.

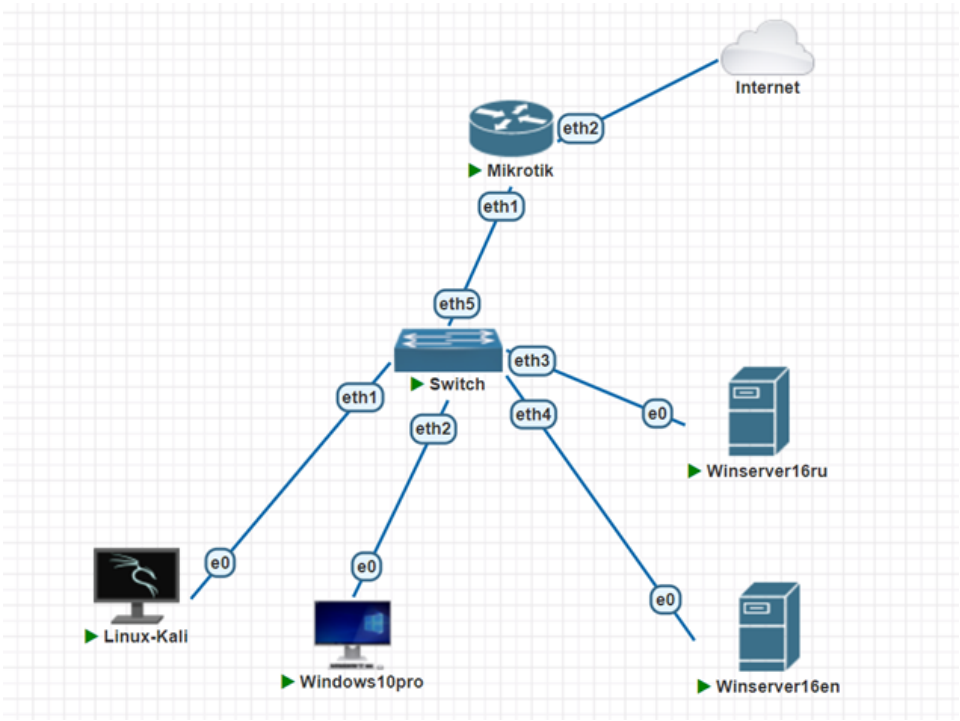


Рисунок 1 – Тестовый стенд

Для выполнения атаки вида Path-the-hash была использована учетная запись и её хэш- пароль, которые был получен ранее. С помощью такой атаки можно выполнять любые действия, доступные пользователю, не имея оригинала пароля.

Crackmapexec – набор python-скриптов для тестирования Windows-окружения, входящий в пакет impacket, использует Windows API, что довольно сложно обнаружить. Базовое применение – быстрое сканирование сети и обнаружение хостов, которые так или иначе взаимодействуют с протоколом smb с возможностью выполнять команды от имени пользователя на удаленной машине посредством командной строки. Также можно использовать утилиту smbexec из пакета impacket, позволяющую запустить cmd windows для удаленной передачи команд. На рисунке 2 показано окно программы с анализом запросов.

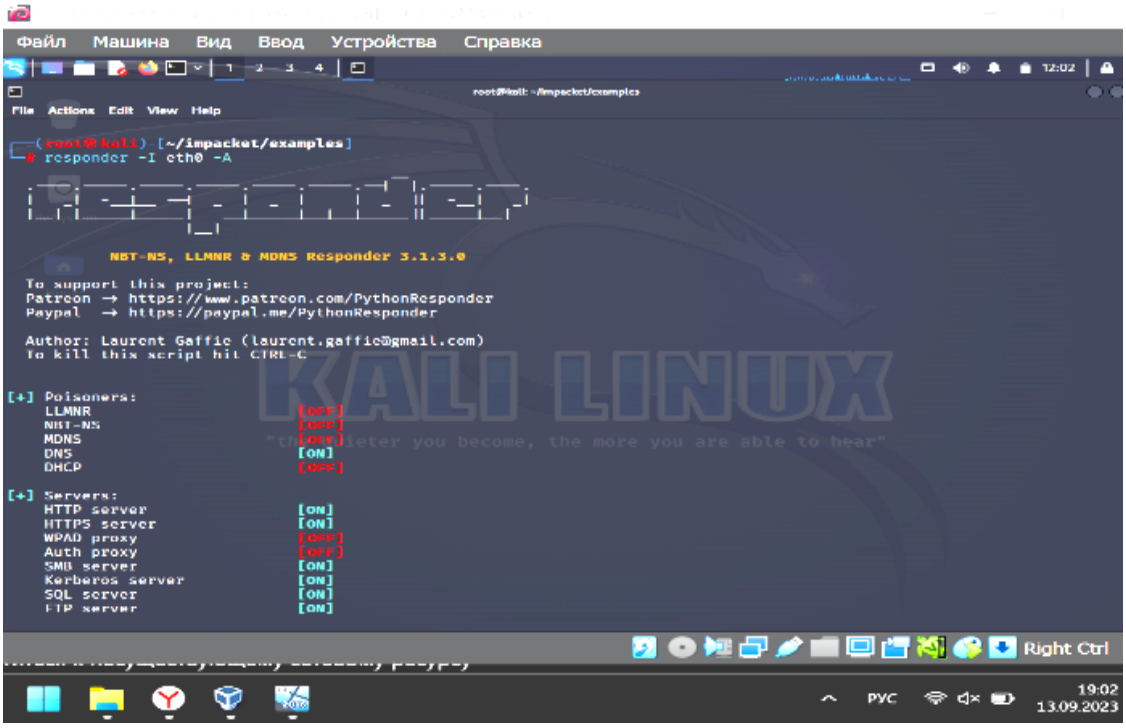


Рисунок 2 – Анализ запросов LLMNR и NBNS с помощью Responder

XFreeRDP – с помощью хеш-пароля пользователей можно зайти в Windows по RDP (протокол удаленного рабочего стола) при условии, что на атакуемом компьютере включен удаленный доступ по RDP. Атаки этого типа будут совершаться с помощью специального протокола WPAD, который используют для прокси-настройки. Основная уязвимость в этих протоколах – это отсутствие подтверждения информации. Таким образом можно сгенерировать любой ответ на запрос ПК по этим протоколам [5].

В ходе работы были применены уязвимости следующих протоколов:

– NBT-NS (NetBIOS Name Service) – протокол, позволяющий компьютерам в локальной сети обращаться друг к другу по именам, не используя сервер DNS;

– LLMNR (Link-Local Multicast Name Resolution) – протокол для поиска и преобразования базовых имен внутри небольшой сети. Использует широковещательные запросы (т. е. такой запрос в сети могут увидеть все);

– mDNS (Multicast DNS) – суть такая же, как и у LLMNR, т. е. DNS-запросы передаются по сети широковещательными запросами. Все вместе технологии реализуют Zeroconf [6].

Основная уязвимость в этих протоколах – отсутствие подтверждения информации. Мы можем сгенерировать любой ответ на запрос ПК по этим протоколам [5].

Для того чтобы автоматически отвечать на подобные запросы, используют инструмент Responder. Атака не пройдёт, если реальный DNS-сервер сможет ответить на DNS запрос компьютера.

В ходе работы был выполнен анализ инфраструктуры через responder. Для этого был запущен анализ приходящих на сетевой интерфейс запросов LLMNR и NBNS с помощью утилиты Responder. Был сгенерирован запрос с PC1, пытаясь обратиться к несуществующему сетевому пути. Например, пользователь попытался пройти по сетевому пути и допустил опечатку, либо предоставил некорректную ссылку.

Далее преднамеренно еще раз был введен неправильный путь для того, чтобы появилось окно авторизации, а Responder – аутентификационный токен, который можно взломать (перебор всех возможных комбинаций символов до тех пор, пока не будет найдена правильная комбинация). В выводе можно увидеть NTLNv2 аутентификацию, которую после можно взломать через такие программы, как hashcat или John the ripper.

В ходе работы была тестирована атака с помощью mitm6. Суть в том, что не был использован протокол ipv6 (который имеет более высокий приоритет, нежели ipv4), отключен его на тестовых серверах DC1 и DC2 и включен на PC1. Несмотря на то, что он не был использован, он все равно работает, следовательно, можно с ним взаимодействовать: имитировать DHCPv6 сервер и отправлять на компьютеры в сети параметры IPv6. Такую атаку можно провести с использованием одноименной утилиты, написанной на Python, – mitm6. Вначале были подменены параметры у PC1, обеспечив mitm. Чтобы атака прошла незаметно для пользователя, нужно воспользоваться способом, который позволит нам мимикрировать под запрашиваемый пользователем ресурс, который требует аутентификацию. Для этого, не отключая mitm6, был создан сервер SMB с помощью программы smbserver.py, а на Win10 через проводник можно зайти на домен и увидеть данные аутентификации в выводе программы.

Уязвимость в протоколе шифрования Zerologon, используемом службой Netlogon, представляет опасность для компьютеров, которые используют этот протокол для аутентификации на контроллере домена и обновления пароля своего аккаунта в Active Directory. Zerologon имеет недостаток в схеме криптографической аутентификации, которую использует Netlogon Remote Protocol (MS-NRPC). Таким образом, злоумышленник может выдавать себя за контроллер домена и изменять пароли пользователей. Получение доступа к контроллеру домена с наивысшими привилегиями дает злоумышленнику полный контроль над корпоративной сетью [6–9].

Суть эксплуатации уязвимости Zerologon:

– с помощью MS-NRPC аутентифицироваться на сервере и сбросить пароль машинной учетной записи;

– атакуется шифрование, возникающее в процессе «рукопожатия» при авторизации для использования команд MS-NRPC. В качестве ключа аутентификации берется строка, состоящая из нулей. А один из 256 векторов инициализации будет состоять из нулей, что приведет к нулевой строке вывода и выполнению MS-RPC команд.

Чтобы эксплуатировать уязвимость Zerologon, был скачен одноименный эксплойт с GitHub (GitHub – это служба размещения в интернете репозиториев Git). После запуска скрипт обнулил пароль машинной учетной записи контроллера домена. Далее была осуществлена распаковка NTDS с помощью secretsdump, входящего в пакет Zerologon, позволяющая эксплуатировать любые уязвимости вида Path-the-hash.

Была осуществлена проверка раздела Security, в ходе которой появилось событие 4742 (которое указывает на то, что учетная запись компьютера была изменена). Внимательно его проанализировав, можно заметить ANONYMOUS LOGON, что является «довольно подозрительным фактом», также в логе можно увидеть заполненное поле password last set – значит, что пароль был изменён. Учитывая предложенные нами настройки системы, менять пароль разрешено только NETWORK SERVICE. При изменении пароля генерируется событие с id 5823. Было найдено событие 5823 в журнале System с помощью фильтра, что показало, что это событие не происходило за этот день (рис. 3).

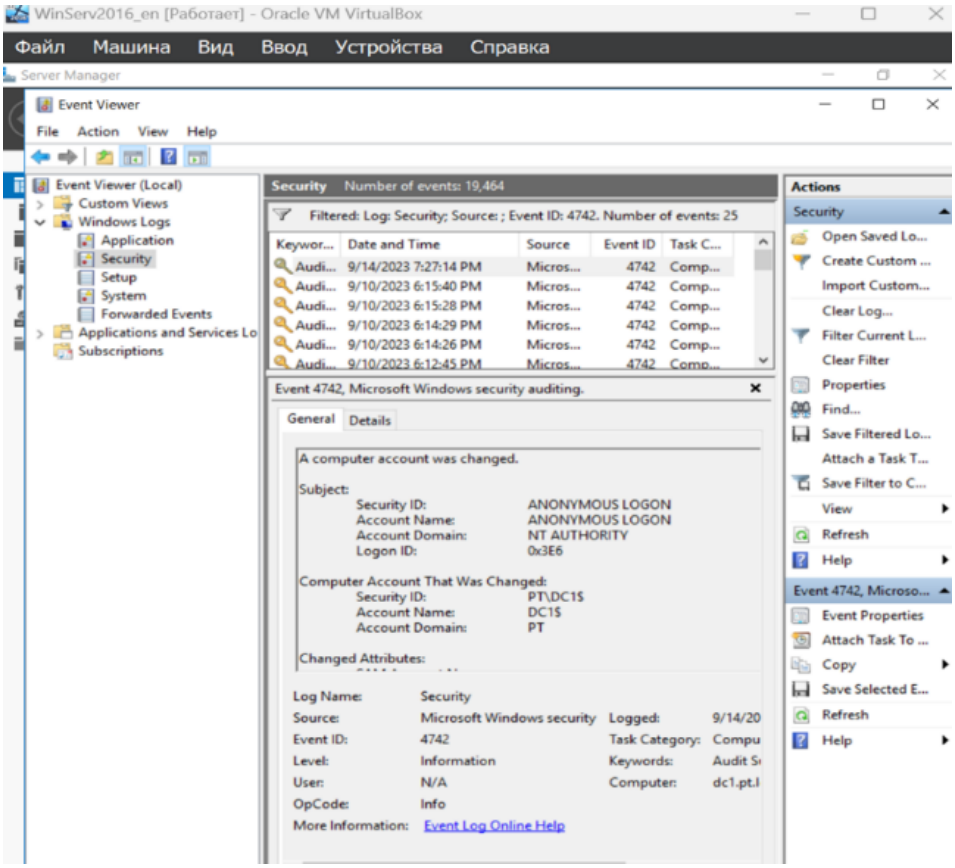


Рисунок 3 – Просмотр события 4742

ЗАКЛЮЧЕНИЕ

Современная информационная инфраструктура подвергается постоянным угрозам, и ее безопасность имеет критическое значение для организаций. Актуальные угрозы включают в себя разнообразные атаки, направленные на сетевые системы и данные. Внимание угрозам, уязвимостям и мерам безопасности на различных платформах, таких как Windows, является необходимым элементом обеспечения безопасности.

Одной из ключевых составляющих безопасности Windows-инфраструктуры является понимание актуальных уязвимостей и их эксплуатация. Системы, работающие на операционной системе Windows, подвержены различным угрозам, и администраторам необходимо постоянно обновлять их и внимательно следить за уязвимостями. Это включает в себя регулярное применение обновлений, а также мониторинг событий и логов для обнаружения угроз.

Одной из возможных угроз для инфраструктуры Windows является эксплуатация контроллера домена. Контроллеры домена играют важную роль в управлении активами и пользователями в сети Windows, и их компрометация может иметь серьезные последствия. Поэтому необходимо уделять особое внимание обеспечению безопасности контроллеров домена и мониторингу активности в сети.

В целом, для обеспечения безопасности информационной инфраструктуры необходимо учитывать актуальные угрозы, следить за уязвимостями, обновлять системы и активно мониторить события в сети. Реакция на угрозы должна быть быстрой и эффективной, чтобы минимизировать потенциальный ущерб для организации. Разработанный тестовый блок на практике показал возможности появления и предотвращения угроз различного вида.

Список источников

1. Лукацкий А. Обнаружение атак / А. Лукацкий. – 2019. – 416 с.
2. Ralf, Hacker. Active Directory глазами хакера / Ralf Hacker. – БХВ-Петербург, 2021. – 176 с.
3. Платунова С. М. Windows Server 2012. Управление серверами. Автоматизация административных задач : учеб. пос. по дисциплине «Администрирование вычислительных сетей». – Санкт-Петербург : НИУ ИТМО, 2016. – 126 с.
4. Дишан, Франсис. Active Directory. Полное Руководство / Дишан Франсис. – 3-е изд. – 2013. – 512 с.
5. Джеймс, Форшоу. Атака сетей на уровне протоколов / Джеймс Форшоу. – 2021. – 342 с.

6. Левицкий Н. Д. Справочник системного администратора. Полное руководство по управлению Windows-сетью / Н. Д. Левицкий. – Санкт-Петербург : Наука и Техника, 2020. – 464 с.
7. Брэгг, Р. Безопасность сетей: полное руководство / Р. Брэгг, М. Родс-Оусли, К. Страссберг. – 2011. – 912 с.
8. Галицкий, А. В. Защита информации в сети – анализ технологий и синтез решений / А. В. Галицкий. – 2010. – 615 с.
9. Уваров, А. Искусство защиты от атак Path-to-Path / А. Уваров. – 2018. – 336 с.

References

1. Lukatsky, A. *Detection of attacks*, 2019. 416 p.
2. Ralf, Hacker. *Active Directory through the eyes of a hacker.* BHV-Petersburg, 2021. 176 p.
3. Platunova, S. M. *Windows Server 2012. Server management. Automation of administrative tasks : textbook for the discipline "Administration of computer networks".* St. Petersburg, National Research University ITMO, 2016. 126 p.
4. Dishan, Francis. *Active Directory. Complete Guide*, 3rd ed., 2013. 512 p.
5. James, Forshaw. *Attacking networks at the protocol level*, 2021. 342 p.
6. Levitsky, N. D. *System Administrator's Handbook. The Complete Guide to Windows Network Management.* St. Petersburg, Science and Technology, 2020. 464 p.
7. Bragg, R., Rhodes-Owsley, M., Strassberg, K. *Network security: a complete guide*, 2011. 912 p.
8. Galitsky, A. V. *Protecting information on the network – analysis of technologies and synthesis of solutions*, 2010. 615 p.
9. Uvarov, A. *The art of protection against Path-to-Path attacks*, 2018. 336 p.

Статья поступила в редакцию 17.10.2023; одобрена после рецензирования 18.08.2023; принята к публикации 20.08.2023.
The article was submitted 17.10.2023; approved after reviewing 18.08.2023; accepted for publication 20.08.2023.

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ, ЧИСЛЕННЫЕ МЕТОДЫ И КОМПЛЕКСЫ ПРОГРАММ

DOI 10.54398/20741707_2023_4_88
УДК 004.896

МОДЕЛИРОВАНИЕ ВЕКТОРА СЕТЕВЫХ АТАК НА ЛОКАЛЬНУЮ СЕТЬ С ПРИМЕНЕНИЕМ БАЗЫ MITRE ATT&CK

Касимова Алина Ринадовна, Казанский национальный исследовательский технологический университет, 420006, Российская Федерация, Казань, ул. Карла Маркса, 68,
старший преподаватель, ORCID: 0000-0001-8927-9113, e-mail: alrkasimova@kstu.ru
Золотарев Вячеслав Владимирович, Сибирский государственный университет науки и технологий, 660037, Российская Федерация, г. Красноярск, пр. им. газ. «Красноярский рабочий», 31,
кандидат технических наук, доцент, заведующий кафедрой безопасности информационных технологий, ORCID: 0000-0002-8054-8564, e-mail: zolotorev@sibsau.ru
Сафиуллина Лина Хатыповна, Казанский национальный исследовательский технологический университет, 420006, Российская Федерация, Казань, ул. Карла Маркса, 68,
кандидат технических наук, доцент, ORCID:0000-0002-2765-0973, e-mail: lina.kh.safiullina@mail.ru
Сабирова Динара Илнуровна, Казанский национальный исследовательский технологический университет, 420006, Российская Федерация, Казань, ул. Карла Маркса, 68,
кандидат химических наук, доцент, ORCID: 0009-0007-5066-5907, e-mail: dinka-sab@mail.ru

Угроза кибератак стала серьезной проблемой для организаций, решение которой в том числе лежит и в плоскости моделирования сценариев возможных атак с применением технологий цифровых двойников. Полученные сценарии содержат тактики и техники, которые могут быть полезны для внедрения и/или корректировки регламента реагирования на инциденты информационной безопасности на предприятии. В представленной статье описан способ проектирования цифрового двойника на базе виртуального стенда автоматизированной информационной системы с указанием конкретных действий злоумышленника в рамках атаки Cyber Kill Chain, когда он, используя различные тактики и техники, прошел все этапы атаки и успешно реализовал свою цель – снятие дампа базы данных. Все этапы подготовки, внедрения и реализации угроз построены на базе знаний MITRE ATT&CK, а также с учетом методик, предлагаемых российскими регуляторами в области защиты информации.

Ключевые слова: АИС, ИБ, MITRE ATT&CK, цифровой двойник, сетевая атака

Финансирование: *исследование выполнено при финансовой поддержке Минцифры РФ (грант ИБ), проект № 40469-01/2022-д.*

MODELING THE NETWORK ATTACK VECTOR ON A LOCAL NETWORK USING THE MITER ATT&CK

Kasimova Alina R., Kazan National Research Technological University, 68 Karl Marks St., Kazan, 420006, Russian Federation,
Senior Lecturer, ORCID: 0000-0001-8927-9113, e-mail: alrkasimova@kstu.ru
Vyacheslav Zolotarev V., Siberian State University of Science and Technology, 31 Krasnoyarsky Rabochy Ave, Krasnoyarsk, 660037, Russian Federation,
Cand. Sci. (Engineering), Assistant Professor, Head of The Information Technologies Security Department, ORCID: 0000-0002-8054-8564, e-mail: zolotorev@sibsau.ru
Safiullna Lina Kh., Kazan National Research Technological University, 68 Karl Marks St., Kazan, 420006, Russian Federation,
Cand. Sci. (Engineering), Assistant Professor, ORCID: 0000-0002-2765-0973, e-mail: lina.kh.safiullina@mail.ru
Sabirova Dinara I., Kazan National Research Technological University, 68 Karl Marks St., Kazan, 420006, Russian Federation,
Cand. Sci. (Chemistry), Assistant Professor, ORCID: 0009-0007-5066-5907, e-mail: dinka-sab@mail.ru

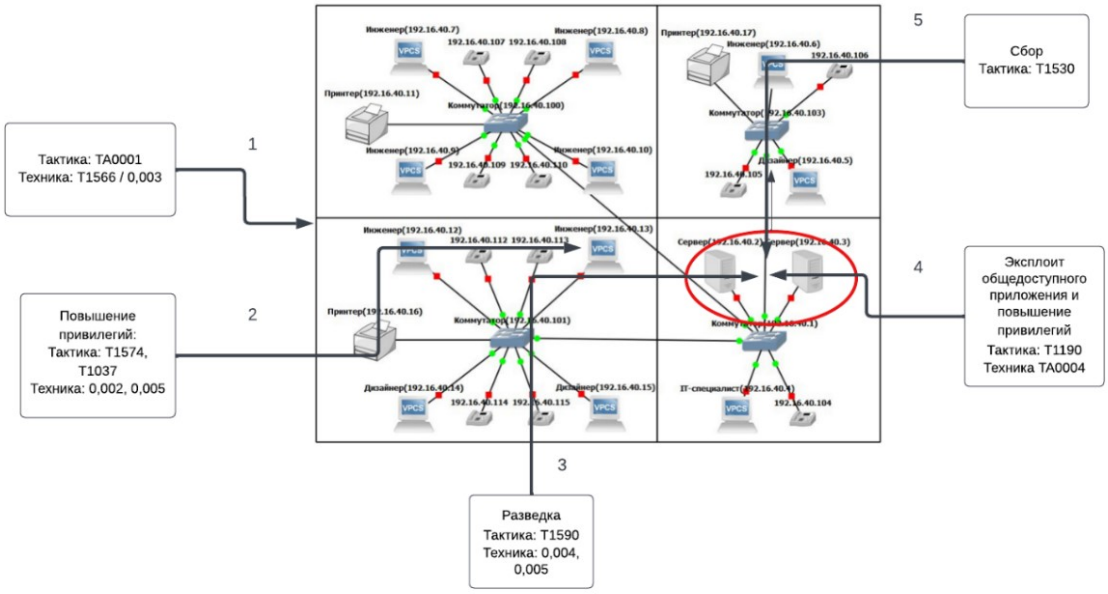
The threat of cyber attacks has become a serious problem for organizations, the solution of which also lies in modeling scenarios of possible attacks using digital twin technologies. The resulting scenarios contain tactics and techniques that can be useful for implementing and/or adjusting the regulations for responding to information security incidents in an enterprise. This article describes a method for designing a digital twin based on a virtual stand of an automated information system, indicating the specific actions of an attacker as part of a Cyber Kill Chain attack,

when he, using various tactics and techniques, went through all stages of the attack and successfully achieved his goal – taking a database dump. All stages of preparation, implementation and implementation of threats are built on the MITER ATT&CK knowledge base, as well as taking into account the methods proposed by Russian regulators in the field of information security.

Keywords: AIS, information security, MITER ATT&CK, digital twin, network attack

Financial support: *the work was supported by the Ministry of Digital Development of the Russian Federation (Grant Information Security), project no. 40469-01/2022-д.*

Graphical annotation (Графическая аннотация)



ВВЕДЕНИЕ

Одними из актуальных вопросов проектирования современных автоматизированных информационных систем (АИС) являются проблемы обеспечения информационной безопасности [1–2]. При этом даже безопасность отдельных подсистем АИС не гарантирует безопасность всей системы в целом. Следовательно, функциональная безопасность АИС в целом требует всестороннего анализа и изучения на каждом этапе обработки информации [3]. Однако существующие методики и технологии проектирования рассматривают АИС без привязки к специфичным процессам конкретного производства, что не позволяет учесть особенности тех или иных функций локальной сети и, как следствие, упустить из поля зрения возможные векторы атак на систему. В то же время при попытке моделирования АИС возникают затруднения с реализацией практической части исследования: построение топологии сети, настройка интерфейсов, взаимодействие сетевых протоколов, эмуляция кибератак и изучение возможных методов противодействия им [4].

В качестве источников информации об инцидентах компьютерной безопасности в АИС рассматриваются такие ее элементы, как: серверы баз данных, компьютерной сети, автоматизированные рабочие места (АРМ), файерволы, IDS/IPS, антивирусное ПО, виртуальные сети, коммутаторы и маршрутизаторы и др., а также прикладное ПО, всевозможные устройства полевого уровня (датчики, микроконтроллеры, исполнительные устройства и пр.) [5–6].

При обнаружении кибератаки на АИС необходимы корректное классифицирование и верные действия по локализации, после окончания кибератаки – устранение уязвимостей и проведение оценки угроз. Для моделирования кибератак и обнаружения возможного вектора угроз и действий нарушителя успешно применяются технологии цифровых двойников [7], которые могут быть использованы для решения широкого круга задач информационной безопасности (ИБ). Интересными в таком случае являются алгоритмы сбора данных, способы безопасного тестирования известных уязвимостей и их влияние на работы АИС, анализ возможности сбора и обработки цифровых свидетельств, обогащения баз данных ИБ и баз знаний инструментальных средств аналитики ИБ. В качестве дополнительной информации о возможных действиях злоумышленника можно использовать MITRE ATT&CK.

MITRE ATT&CK® – это общедоступная база знаний о тактике и методах противника, основанная на реальных наблюдениях. База знаний ATT&CK используется в качестве основы для разработки конкретных моделей угроз и методологий в частном секторе, правительстве, а также

в сообществе продуктов и услуг кибербезопасности. Создавая ATT&CK, MITRE выполняет свою миссию по решению проблем для более безопасного мира, объединяя сообщества для разработки более эффективной кибербезопасности. ATT&CK открыт и доступен любому лицу или организации для бесплатного использования. В базе знаний MITRE ATT&CK содержатся техники и технологии, использующиеся при большинстве распространенных типов атак. Следовательно, основывая модель информационной безопасности АИС на MITRE ATT&CK, можно правильно выстроить стратегию защиты и успешно провести дальнейшее расследование компьютерного инцидента. Успешный пример применения базы знаний MITRE ATT&CK при проектировании системы управления судоходными средствами представлен в [8].

В настоящей статье предложен метод моделирования вектора сетевых атак на АИС с применением базы MITRE ATT&CK с предсказанием возможных действий злоумышленника, приведены результаты пробного запуска сценария подобного типа атак, а также представлены возможности вероятного использования полученных данных [9].

ОПИСАНИЕ ВИРТУАЛЬНОЙ ЛАБОРАТОРИИ НА БАЗЕ ЦИФРОВОГО ДВОЙНИКА

Цифровой двойник, используемый как тестовый стенд, должен иметь безопасное соединение как с устройствами управления основным технологическим процессом как в случае обмена данными в реальном времени, так и в случае обмена данными в асинхронном режиме. Во-вторых, цифровой след, собираемый при работе с цифровым двойником, должен быть полезен для решения задач управления информационной безопасностью реального объекта.

Следовательно, последовательность шагов по интеграции цифрового двойника в качестве лабораторного стенда может быть следующей:

1. Оценка возможных источников данных для формирования цифрового следа.
2. Передача данных как элемент задачи управления безопасностью должна содержать формат самих данных, формат заголовков и метаданных, стандартные формы и протоколы, отчеты и поля баз данных, используемые для хранения.
3. Так как данные цифрового следа будут храниться как в базах данных реального объекта, так и в базах данных цифрового двойника, необходимо продумать способы хранения (архивирования) данных, в том числе технологических параметров и аварийной сигнализации. В число способов защиты данных должны входить контроль доступа, резервирование, управление ключами и коммуникациями и другие применимые в задаче средства управления безопасностью (контрмеры).
4. С точки зрения использования данных должны быть проанализированы вопросы доступа к данным, в том числе и с позиции использования автоматического доступа средствами оркестрации в системе управления информационной безопасностью.

Таким образом, при использовании цифрового двойника реального объекта возникает положительная обратная связь с процедурой сбора цифровых свидетельств событий информационной безопасности, реализованной в цифровом двойнике (и/или реальном объекте) и процедурой повышения осведомленности специалистов по защите информации, реализующих задачи управления информационной безопасности в рамках работы с реальным объектом.

Рассмотрим тестовый стенд сети корпорации. Корпорация занимается дизайном и разработкой определенных инженерных и дизайнерских решений для более крупных компаний. На сайте компании есть личный кабинет клиента для более удобного взаимодействия с клиентами. Также имеется список всех работников с рабочими мессенджерами для связи (рабочие сообщения в них просматриваются руководителем, основная их цель – создание атмосферы более тесного общения с будущими или действующими клиентами). Также на сайте представлены фото офиса с АРМ, на рабочих столах которых видны программы, которыми пользуется компания (PS C5 Adobe, 3D Компас, Mac Apps, Office 2010). На рисунке 1 представлены схемы сети:

IP-телефония: 11 шт.;

АРМ инженера – ОС Windows 10: 6 шт.;

АРМ дизайнера – Mac OS: 2 шт.;

АРМ секретаря – ОС Windows 8.1 – 1 шт.;

АРМ IT-специалиста – ОС Windows 10, Kali Linux – 1 шт.;

сервер 1 (192.16.40.2) – Debian 10 server – (сайт компании, база данных MySQL – PhpMyAdmin);

сервер 2 (192.16.40.3) – Ubuntu Server – (сервер Телефонии, сервер Печати, база Данных MySQL – PhpMyAdmin, локальное облачное хранилище, почтовый сервер, VipNet);

роутер – Microtik – 1 шт.;

лазерный принтер – Linux Type OS – 2 шт. (HP LaserJet Professional M1210 MFP) ;

коммутатор – S2700 – 52P-EI-AC коммутатор Huawei – 2 шт.

Объектами защиты могут выступать [10]:

1. Оценка снимаемых параметров, их формата и способа считывания для анализа возможности применения защитных мер.

2. Анализ протоколов обмена данными для получения информации о способе передачи, формате данных и заголовков, служебной информации, промежуточных коммуникационных устройствах.

3. Анализ системы управления хранением данных и их обработкой. В большинстве случаев речь будет идти либо о работе с файлами, в том числе большого размера, либо о системе управления базами данных. Соответственно, меры по защите информации будут сосредоточены либо на безопасности штатных средств обработки данных, либо на защите учетных записей, привязанных к их обработке. Также допустимо сквозное шифрование, в том числе «легковесными» криптоалгоритмами.

4. Анализ порядка доступа к данным. Это может быть значимо как для управления технологическим процессом, так и для извлечения и обработки данных. Имеет значение используемая операционная среда, а также возможности штатных средств доступа, такие как резервное копирование и защита информации резервных копий, идентификация и аутентификация, шифрование данных.

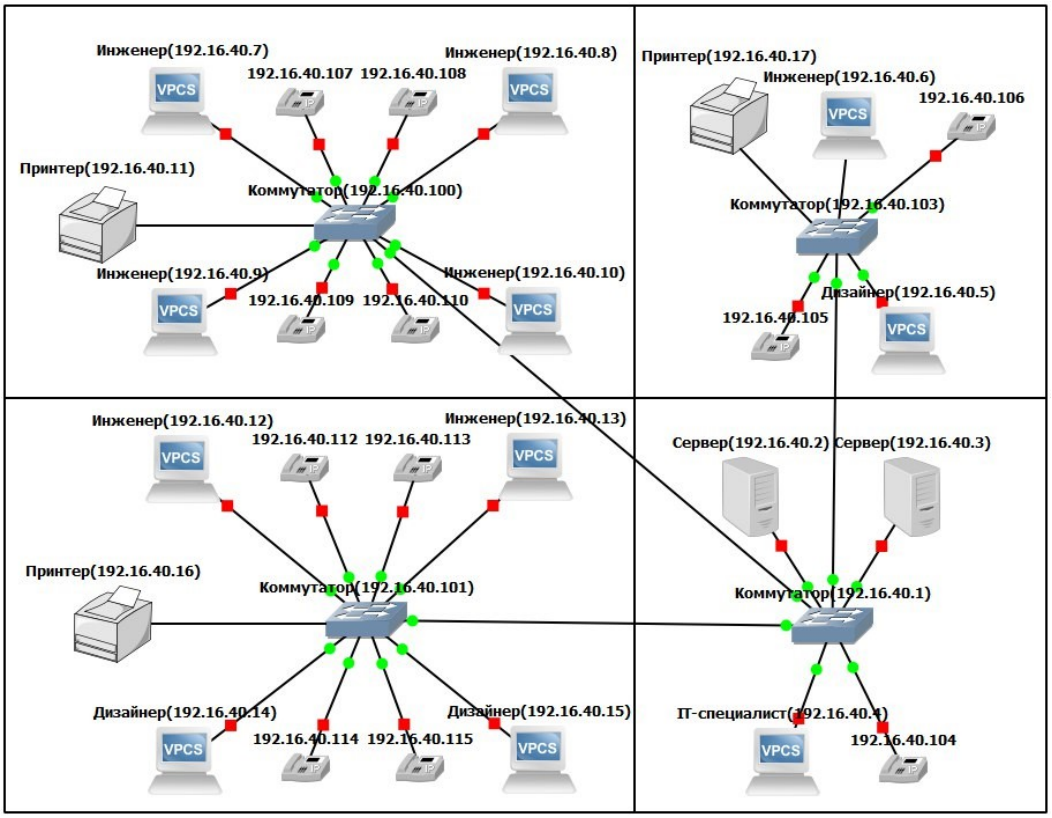


Рисунок 1 – Офисная сеть корпорации

ПРЕДСКАЗАНИЕ ВЕКТОРА ДЕЙСТВИЙ ЗЛОУМЫШЛЕННИКА С ПРИМЕНЕНИЕМ БАЗЫ MITRE ATT&CK

Согласно методике оценки угроз безопасности информации ФСТЭК [11] можно выделить следующие виды и цели внешнего нарушителя, актуальные для предприятия (табл.).

Таблица – Виды и цели внешнего нарушителя

Вид нарушителя	Возможные цели реализации
Преступные группы (криминальные структуры)	Получение финансовой или иной материальной выгоды. Желание самореализации (подтверждение статуса)
Отдельные физические лица (хакеры)	Получение финансовой или иной материальной выгоды. Любопытство или желание самореализации (подтверждение статуса)
Конкурирующие организации	Получение конкурентных преимуществ. Получение финансовой или иной материальной выгоды
Лица, обеспечивающие поставку программных, программно-аппаратных средств, обеспечивающих систем	Получение финансовой или иной материальной выгоды. Непреднамеренные, неосторожные или неквалифицированные действия. Получение конкурентных преимуществ

Согласно методике ФСТЭК, нарушитель обладает средними возможностями, а именно:

1) приобретать информацию об уязвимостях, размещаемую на специализированных платных ресурсах (биржах уязвимостей);

2) приобретать дорогостоящие средства и инструменты для реализации угроз, размещаемые на специализированных платных ресурсах;

3) получать доступ к встраиваемому программному обеспечению аппаратных платформ, системному и прикладному программному обеспечению, телекоммуникационному оборудованию и другим программно-аппаратным средствам для проведения их анализа;

4) обладает знаниями и практическими навыками проведения анализа программного кода для получения информации об уязвимостях;

5) обладает высокими знаниями и практическими навыками о функционировании систем и сетей, операционных систем, а также имеет глубокое понимание защитных механизмов, применяемых в программном обеспечении, программно-аппаратных средствах;

6) реализовывать угрозы безопасности информации в составе группы лиц.

Злоумышленник использует вектор атаки и набор TTP (техники, тактики и процедуры), которые расшифровываются следующим образом [12]:

1) тактика – как злоумышленник действует на разных этапах своей операции, какая цель или задача злоумышленника на определенным шаге;

2) техника – как злоумышленник достигает цели или поставленной задачи, какие использует инструменты, технологии, код, эксплоиты, утилиты и т. д.;

3) процедура – как эта техника выполняется и для чего.

Всего в матрице MITRE ATT&CK [13] 14 тактических задач злоумышленника при атаке на организацию, а именно:

1. Рекогносцировка (начальный доступ).
2. Разработка ресурсов.
3. Первоначальный доступ.
4. Внедрение и исполнение вредоносного кода.
5. Закрепление.
6. Повышение привилегий.
7. Обход защиты.
8. Получения учетных данных.
9. Развертывание.
10. Боковое перемещение.
11. Сбор данных.
12. Управление и контроль.
13. Эксфильтрация.
14. Воздействие.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Возможная цель атаки злоумышленника: получение базы данных клиентов компании.

Смежная цель: создать прямой VPN-туннель для получения информации, шантажа, слива информации.

Начальный доступ

Тактика: TA0001.

Техника: T1566 / 0,003 Фишинг/ субтехника Целевой фишинг через сервис.

Описание: злоумышленник отправляет правление на почту корпорации и в личные сообщения (адреса и номера есть на сайте компании) сообщение с зараженным файлом для повышения уровня доступа относительно локальной сети.

Пример сообщения

<Здравствуйте, наша компания хотела бы заказать разработку концепта для умных бытовых устройств, пожалуйста, рассмотрите гиф с нашими концептами, которые продаются, и схемами. Можете достать их из GIF, прикрепленного ниже.>

Злоумышленник может отправлять фишинговые сообщения через сторонние службы в попытке получить доступ к системам жертв. Спирфишинг через сервис является специфическим вариантом целевого фишинга. Он отличается от других форм целевого фишинга тем, что он использует сторонние сервисы, а не непосредственно корпоративные каналы электронной почты.

Задача нарушителя в ходе стандартной работы в виде странного заказчика дойти до того, чтобы инженер открыл файл со схемами в программе Photoshop, запустив тем самым вредоносный скрипт, который может вводить производный код. На рисунке 2 показан пример вредоносного файла, инкапсулированного через ОС Kali Linux.

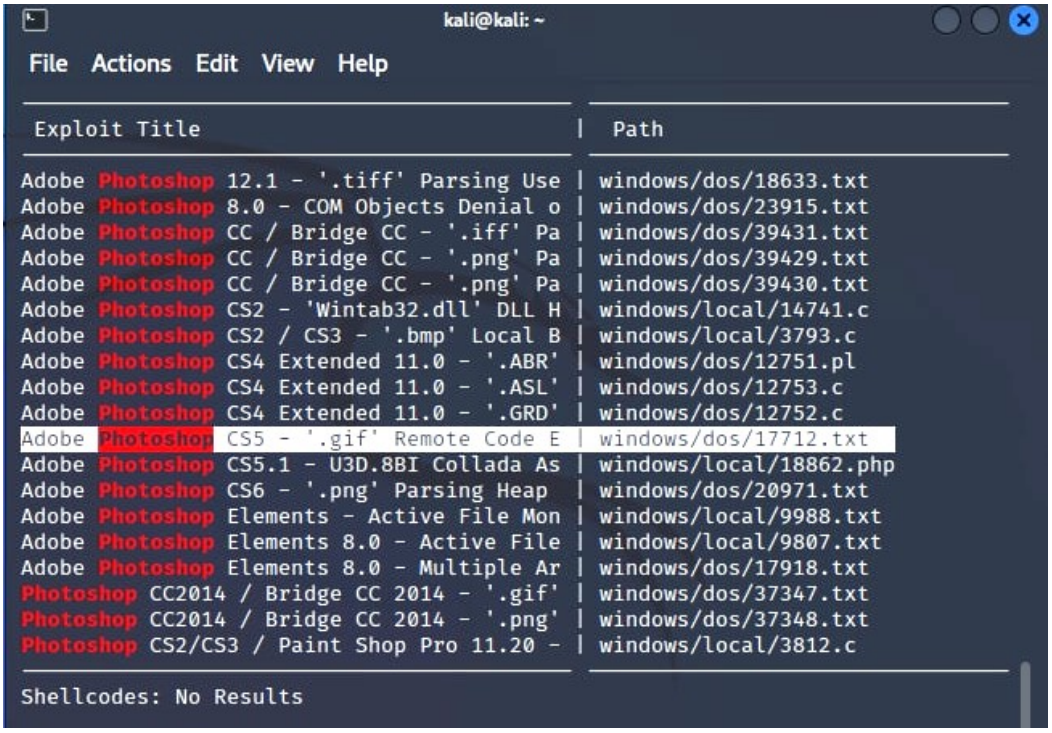


Рисунок 2 – Пример вредоносного файла

Повышение привилегий
Тактика: T1574, T1037.
Техника: 0,002, 0,005.
Описание:

После открытия файла был запущен код, который провел установку и запуск NJRat. Таким образом, нарушитель получил полный удаленный доступ к рабочей станции, а также всю информацию о подключениях. На рисунке 3 показана информация хоста 192.16.40.13.

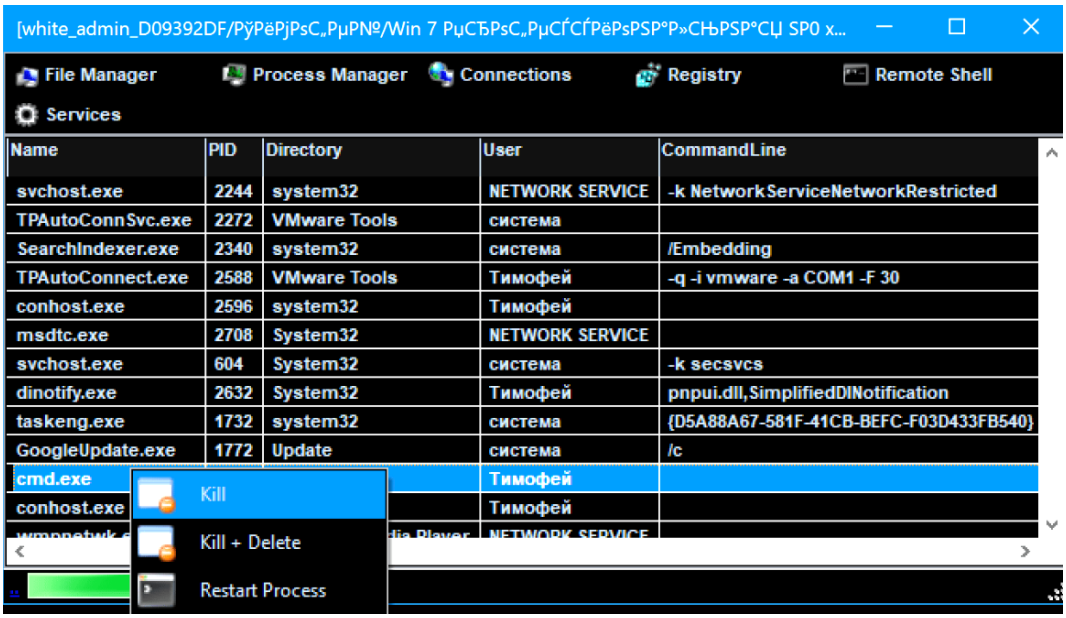


Рисунок 3 – Информация из зараженного хоста, полученная NJRat

Разведка
Тактика: T1590.
Техника: 0,004, 0,005.

Описание:

Злоумышленники могут собирать информацию о топологии сети жертвы, которую можно использовать во время нацеливания. Информация о сетевых топологиях может включать множество деталей, в том числе физическое и/или логическое расположение как внешней, так и внутренней сетевой среды. Эта информация может также включать сведения о сетевых устройствах (шлюзах, маршрутизаторах и т. д.) и другой инфраструктуре.

После определения операционной системы и портов можно понять, какие клиенты сети являются серверами.

Запускается сканирование уязвимостей через OpenVas.Tenable Nessus.Gobysec / Goby.Tsunami-security-scanner. Flan Scan. D9scan. Rustscan.

Сканер запущен на 192.16.40.2 и 192.16.40.3. Результат работы сканера представлен на рисунке 4.

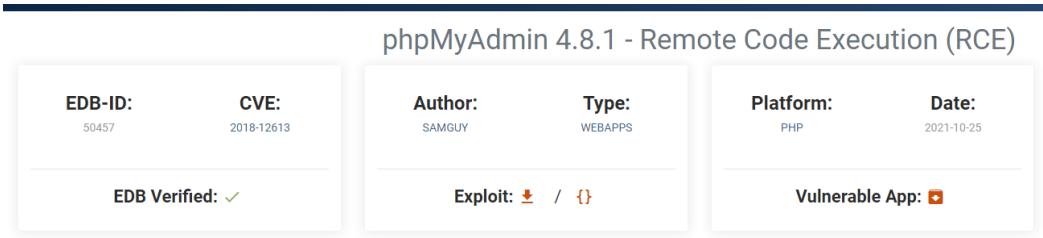


Рисунок 4 – Вывод сканера на 192.16.40.2, 192.16.40.3 с уязвимостью CVE-2018-12613

У серверов имеется ошибка, позволяющая запускать произвольный код.

Эксплоит общедоступного приложения и повышение привилегий

Тактика: T1190

Техника TA0004

Описание:

Злоумышленники могут попытаться воспользоваться слабостью компьютера или программы, подключенной к Интернету, используя программное обеспечение, данные или команды, чтобы вызвать непреднамеренное или непредвиденное поведение. Слабостью системы может быть ошибка, глюк или уязвимость дизайна. Эти приложения часто являются веб-сайтами, но могут включать базы данных (например, SQL), стандартные службы (такие как SMB или SSH), протоколы администрирования и управления сетевыми устройствами (такие как SNMP и Smart Install) и любые другие приложения с открытыми сокетами, доступными через Интернет, такие как веб-серверы и связанные службы.

В сценарии запуск MetaSploitFrame. Rhost для атаки по очереди отправляет произвольный код, который будет прокинут связан с созданием пользователя с sudo правами, и ssh доступом, и созданием администратора в myphpadmin на 192.16.40.2, 192.16.40.3 на порт 80 и 3306.

Сбор

Тактика: T1530

Описание:

Злоумышленники могут получить доступ к данным из ненадлежащим образом защищенного облачного хранилища. Злоумышленники могут собирать конфиденциальные данные из этих облачных хранилищ. Поставщики обычно предлагают руководства по безопасности, чтобы помочь конечным пользователям настроить системы, хотя неправильная конфигурация является распространенной проблемой. Имели место многочисленные инциденты, когда облачное хранилище было ненадлежащим образом защищено, как правило, путем непреднамеренного предоставления общего доступа пользователям, не прошедшим проверку подлинности, чрезмерно широкого доступа для всех пользователей или даже доступа для любого анонимного лица, находящегося вне контроля системы управления доступом к учетным записям, даже не требуя базовых разрешений пользователя.

Этот открытый доступ может раскрывать различные типы конфиденциальных данных, таких как кредитные карты, личная информация или медицинские записи.

Злоумышленники также могут получить, а затем злоупотребить утечкой учетных данных из исходных репозиториев, журналов или других средств, чтобы получить доступ к объектам облачного хранилища. Конечная цель злоумышленника – дамп базы данных – достигнута.

ЗАКЛЮЧЕНИЕ

В работе была предпринята попытка смоделировать вектор атаки злоумышленника (Cyber Kill Chain), используя базу знаний MITRE ATT&CK, а также методики, предлагаемые российскими регуляторами в области защиты информации.

В качестве тестового стенда был взят цифровой двойник АИС предприятия для задач управления изменениями и применения данных о функционировании системы защиты информации в режимах противодействия атаке и минимизации ущерба, управления инцидентами.

Внешний злоумышленник, используя различные тактики и техники, прошёл все этапы атаки и успешно реализовал свою цель – снятие дампа базы данных. В работе была показана лишь часть процедур, которыми может воспользоваться злоумышленник, для выполнения поставленной цели.

В свою очередь, считаем, что поставленная перед авторами цель работы – рассмотреть потенциально опасный сценарий кибератаки (вектор атаки, Cyber Kill Chain) для предложенного цифрового двойника информационной системы предприятия – успешно достигнута.

Список источников

1. Анацкая, А. Г. Моделирование сценариев атак для центра обработки данных в составе автоматизированной системы / А. Г. Анацкая, А. А. Берендеев // Образование. Транспорт. Инновации. Строительство : сборник материалов V Национальной научно-практической конференции, Омск, 28–29 апреля 2022 года. – Омск : Сибирский государственный автомобильно-дорожный университет (СибАДИ), 2022. – С. 639–644. – EDN FKEIJU.

2. Шевцов, В. Ю. Применение сетей Петри при моделировании атак на системы АСУ ТП / В. Ю. Шевцов, Д. И. Правиков // Вестник современных цифровых технологий. – 2022. – № 13. – С. 32–37. – EDN RYUSGM.

3. Набиуллин, Д. И. Кибербезопасность электрических станций / Д. И. Набиуллин, Р. Р. Вилданов // Методические вопросы исследования надежности больших систем энергетики : 92-е заседание семинара, учрежденного при ИСЭМ СО РАН : в 3-х книгах, Казань, 21–26 сентября 2020 года. – Иркутск : Федеральное государственное бюджетное учреждение науки Институт систем энергетики им. Л.А. Мелентьева Сибирского отделения Российской академии наук, 2020. – Т. 71, кн. 2. – С. 251–254. – EDN DJVJZS.

4. Кадан, А. М. Моделирование сетевых атак в среде Графического Сетевого эмулятора GNS3 / А. М. Кадан, Н. В. Белоголовая // Дистанционные образовательные технологии : материалы VII Международной научно-практической конференции, Ялта, 20–22 сентября 2022 года. – Симферополь : Общество с ограниченной ответственностью «Издательство Типография «Ариал», 2022. – С. 253–258. – EDN VMJMBG.

5. Котенко, И. В. Построение системы интеллектуальных сервисов для защиты информации в условиях кибернетического противоборства / И. В. Котенко, И. Б. Саенко // Труды СПИИРАН. – 2012. – № 3 (22). – С. 84–100. – EDN PCCYKV.

6. Котенко, И. В. Применение технологии управления информацией и событиями безопасности для защиты информации в критически важных инфраструктурах / И. В. Котенко, И. Б. Саенко, О. В. Полубелова, А. А. Чечулин // Труды СПИИРАН. – 2012. – № 1(20). – С. 27–56. – EDN PCCXYX.

7. Касимова, А. Р. Использование цифровых двойников при построении системы безопасности предприятия / А. Р. Касимова, Л. Х. Сафиуллина // Международный форум Kazan Digital Week-2022 : сборник материалов Международного форума, Казань, 21–24 сентября 2022 года / под общ. ред. Р. Н. Минниханова. – Казань : Научный центр безопасности жизнедеятельности, 2022. – С. 291–298. – EDN ODYEKU.

8. Jo, Y. Cyberattack Models for Ship Equipment based on the MITRE ATT&CK Framework / Y. Jo, O. Choi, J. You, Y. Cha, D. H. Lee // Sensors. – 2022. – № 22. – P. 1860. https://doi.org/10.3390/s22051860.

9. Щербина, К. А. Возможности применения матрицы MITRE ATT&CK / К. А. Щербина, М. А. Копашенко, И. С. Поздняк // XXIX Российская научно-техническая конференция : материалы XXIX Российской научно-технической конференции профессорско-преподавательского состава, научных сотрудников и аспирантов университета с приглашением ведущих ученых и специалистов родственных вузов и организаций, Самара, 22–25 марта 2022 года. – Самара : Поволжский государственный университет телекоммуникаций и информатики, 2022. – С. 56–57. – EDN LKKQNC.

10. Zolotarev, V. An Approach to the Implementation of Nonparametric Algorithms for Controlling Multidimensional // Processes in a Production Problem / V. Zolotarev, M. Lapina, D. Liksonova // Lecture Notes in Networks and Systems. – 2023. – 702 LNNS. – P. 250–257.

11. Методический документ методика оценки угроз безопасности информации. – URL: fstec.ru /tekhnicheskaya-zashchita-informatsii/dokumenty/114-spetsialnye-normativnye-dokumenty/2170-metodicheskij-dokument-utverzhden-fstek-rossii-5-fevralya-2021-g.

12. Матрица ATT&CK. Как устроен язык описания угроз и как его используют. – URL: www.xakep.ru/2021/03/17/mitre-att-ck/.

13. Матрица ATT&CK. – URL: www.attack.mitre.org/.

References

1. Anatskaya, A. G., Berendeev, A. A. Modeling attack scenarios for a data processing center as part of an automated system. *Education. Transport. Innovation. Construction: Collection of materials of the V National Scientific and Practical Conference, Omsk, April 28–29, 2022.* Omsk, Siberian State Automobile and Highway University (SibADI), 2022, pp. 639–644. EDN FKEIJU.

2. Shevtsov, V. Yu. Pravikov, D. I. Application of Petri nets in modeling attacks on automated process control systems. *Bulletin of Modern Digital Technologies*, 2022, no. 13, pp. 32–37. EDN RYUSGM.

3. Nabiullin, D. I., Vildanov, R. R. Cybersecurity of power plants. *Methodological issues in studying the reliability of large energy systems: 92nd meeting of the seminar established at the ISEM SB RAS : in 3 books, Kazan, September*

21–26, 2020. Irkutsk, Federal State Budgetary Institution of Science Institute of Energy Systems named after L.A. Melentyev Siberian Branch of the Russian Academy of Sciences, 2020, vol. 71, book 2, pp. 251–254. EDN DJVJZS.

4. Kadan, A. M., Belogolovaya, N. V. Simulation of network attacks in the environment of the GNS3 Graphical Network Emulator. *Distance educational technologies : materials of the VII international scientific and practical conference, Yalta, September 20–22, 2022 of the year*. Simferopol, Limited Liability Company "Publishing House Typography "Arial", 2022, pp. 253–258. EDN VMJMBG.

5. Kotenko, I. V., Saenko, I. B. Construction of a system of intelligent services for information protection in conditions of cyber warfare. *Proceedings of SPIIRAS*, 2012, no. 3 (22), pp. 84–100. EDN PCCYKV.

6. Kotenko, I. V., Saenko, I. B., Polubelova, O. V., Chechulin, A. A. Application of technology for managing information and security events to protect information in critical infrastructures. *Proceedings of SPIIRAS*, 2012, no. 1 (20), pp. 27–56. EDN PCCXYX.

7. Kasimova, A. R., Safiullina, L. Kh. Using digital twins in building an enterprise security system. *International Forum Kazan Digital Week-2022 : collection of materials of the International Forum, Kazan, September 21–24, 2022 of the year ; under the general editorship of R. N. Minnikhanov.* Kazan, Scientific Center for Life Safety, 2022, pp. 291–298. EDN ODYEKU.

8. Jo, Y., Choi, O., You, J., Cha, Y., Lee, D. H. Cyberattack Models for Ship Equipment based on the MITRE ATT&CK Framework. *Sensors*, 2022, no. 22, p. 1860. https://doi.org/10.3390/s22051860.

9. Shcherbina, K. A., Kopashenko, M. A., Pozdnyak, I. S. Possibilities of using the MITER ATT&CK matrix. *XXIX Russian Scientific and Technical Conference : materials of the XXIX Russian Scientific and Technical Conference of Faculty, researchers and graduate students of the university with the invitation of leading scientists and specialists from related universities and organizations, Samara, March 22–25, 2022.* Samara, Volga State University of Telecommunications and Informatics, 2022, pp. 56–57. EDN LKKQNC.

10. Zolotarev, V., Lapina, M., Liksonova, D. An Approach to the Implementation of Nonparametric Algorithms for Controlling Multidimensional Processes in a Production Problem. *Lecture Notes in Networks and Systems*, 2023, 702 LNNS, pp. 250–257.

11. *Methodological document methodology for assessing threats to information security.* Available at: fstec.ru/tekhnicheskaya-zashchita-informatsii/dokumenty/114-spetsialnye-normativnye-dokumenty/2170-metodicheskij-dokument-utverzhden-fstek-rossii-5-fevralya -2021-g.

12. *ATT&CK Matrix. How the threat description language works and how it is used.* Available at: www.xakep.ru/2021/03/17/mitre-att-ck/.

13. *ATT&CK Matrix.* Available at: www.attack.mitre.org/.

Статья поступила в редакцию 02.10.2023; одобрена после рецензирования 09.10.2023; принята к публикации 18.10.2023.

The article was submitted 02.10.2023; approved after reviewing 09.10.2023; accepted for publication 18.10.2023.

DOI 10.54398/20741707_2023_4_97
УДК 519.178/.245:004.94

ИСПРАВЛЕНИЕ АЛГОРИТМА ЗАЛИВКИ ДЛЯ НАХОЖДЕНИЯ ГЕОМЕТРИЧЕСКОГО ОСТОВА В ЗАДАЧАХ ПЕРКОЛЯЦИИ УЗЛОВ НА КВАДРАТНЫХ РЕШЕТКАХ

Гордеев Иван Иванович, Астраханский государственный университет, 414056, Российская Федерация, г. Астрахань, ул. Татищева, 20а,
кандидат физико-математических наук, ORCID: 0000-0001-5036-4791, e-mail: g2i@mail.ru
Касаткин Александр Сергеевич, Московский авиационный институт, 125993, Российская Федерация, г. Москва, Волоколамское шоссе, 4,
магистрант, ORCID: orcid.org/0009-0009-1240-0590, e-mail: sashunya.kasatkin@bk.ru
Овчаренко Сергей Сергеевич, АО «Концерн «Моринформсистема-Агат», 105275, Российская Федерация, г. Москва, Шоссе Энтузиастов, 29,
программист, ORCID: 0009-0009-2373-5353, e-mail: obchapehko@mail.ru
Саенко Наталья Сергеевна, филиал ООО «Газпром информ» в г. Астрахань, 414057, Российская Федерация, г. Астрахань, ул. Николая Островского, 130 лит. А,
инженер-программист 2 категории, ORCID: 0000-0002-3595-6892, e-mail: saenko.natasha@mail.ru

В данной статье сравниваются различные версии алгоритма заливки для нахождения геометрического остова на квадратных решетках. Подробно рассматриваются случаи некорректного определения остова существовавшими ранее версиями алгоритма. Описываются конфигурации барьеров из сочленяющих ячеек, при наличии которых происходит некорректное определение остова, и предлагается способ исправления ошибки в алгоритме. Для исправления алгоритма предлагается использовать понятие ориентации сочленяющей ячейки вместо использовавшегося ранее понятия контактных пар сочленяющей ячейки. Сравниваются результаты работы исправленного алгоритма заливки для определения остова с некорректной версией алгоритма заливки, а также с алгоритмом Грассбергера и алгоритмом Ахунжанова для нахождения геометрического остова на квадратных решетках. Экспериментальное сравнение алгоритмов проведено на квадратных решетках с размерами $L = 5, 7, 10, 14, 20, 28, \ldots, 1280, 1792, 2560, 3584, 5120$. Для каждого размера решетки рассматривалось по 200 случайных заполнений. В результате сравнения обнаружено, что исправленный алгоритм заливки дает точное определение остова, совпадающее с алгоритмом Ахунжанова, а алгоритм заливки с ошибкой дает на больших решетках среднюю долю отличий остова более 35 %.

Ключевые слова: идентификация остова, перколяция узлов, двумерная решетка, открытые граничные условия, алгоритмы на графах, алгоритм Грассбергера, алгоритм Ахунжанова, алгоритм заливки

Благодарности. *Авторы благодарят Романа Тробека (Roman Trobec) из Института Йозефа Стефана в Словении за предоставленную возможность познакомиться с исходными кодами на языке Matlab. Также авторы благодарят Инь Вэйго (Weiguo Yin) из Брукхейвенской национальной лаборатории в США и Тао Жуйбао (Ruibao Tao) из Фуданьского университета в КНР за предоставленную возможность познакомиться с исходными кодами на языке Fortran.*

CORRECTION OF THE FLOODING ALGORITHM FOR FINDING THE GEOMETRICAL BACKBONE IN PROBLEMS OF SITE PERCOLATION ON SQUARE LATTICES

Gordeev Ivan I., Astrakhan Tatishchev State University, 20a Tatishchev St., Astrakhan, 414056, Russian Federation,
Cand. Sci. (Physics and Mathematics), ORCID: 0000-0001-5036-4791, e-mail: g2i@mail.ru
Kasatkin Alexander S., Moscow Aviation Institute, 4 Volokolamskoye Shosse, Moscow, 125993, Russian Federation,
master's student, ORCID: 0009-0009-1240-0590, e-mail: sashunya.kasatkin@bk.ru
Ovcharenko Sergey S., JSC Concern Morinformsystem-Agat, Shosse Entuziastov, 29, Moscow, 105275, Russian Federation,
programmer, ORCID: 0009-0009-2373-5353, e-mail: obchapehko@mail.ru
Saenko Natalya S., Branch of Gazprom Inform LLC in Astrakhan, 130 lit. A Nikolay Ostrovsky St., Astrakhan, 414057, Russian Federation,
software engineer of category 2, ORCID: 0000-0002-3595-6892, e-mail: saenko.natasha@mail.ru

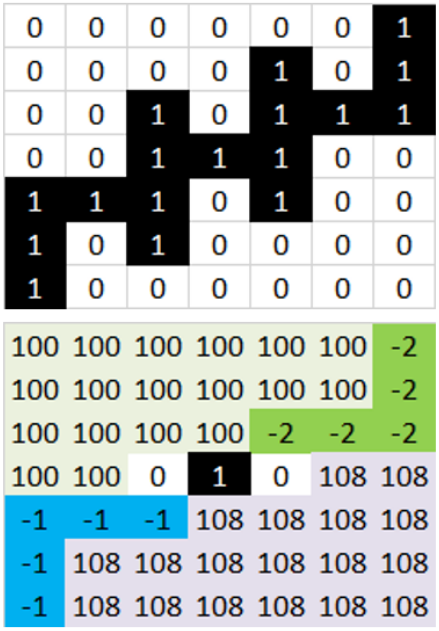
This paper compares different versions of the fill algorithm for finding the geometric backbone on square lattices. Cases of incorrect determination of the backbone by previously existing versions of the algorithm are discussed in detail. The configurations of barriers formed by articulation cells, in the presence of which the backbone is incorrectly determined, are described, and a method for correcting the error in the algorithm is proposed. To correct the algorithm, it is proposed to use the concept of orientation of an articulating cell instead of the previously used concept

of contact pairs of an articulating cell. The results of the corrected flooding algorithm for determining the backbone are compared with the incorrect version of the flooding algorithm, as well as with the Grassberger algorithm and the Akhunzhanov algorithm for finding the geometric backbone on square lattices. An experimental comparison of the algorithms was carried out on square lattices with sizes $L = 5, 7, 10, 14, 20, 28, \dots, 1280, 1792, 2560, 3584, 5120$. For each lattice size, 200 random fillings were considered. As a result of the comparison, it was found that the corrected filling algorithm gives an accurate definition of the backbone, coinciding with Akhunzhanov's algorithm, and the filling algorithm with an error gives on large lattices an average fraction of differences in the backbone of more than 35 %.

Keywords: backbone identification, site percolation, two-dimensional lattice, open boundary conditions, graph algorithms, Grassberger algorithm, Akhunzhanov algorithm, flooding algorithm

Acknowledgments. *The authors thank Roman Trobec from the Josef Stefan Institute in Slovenia for the opportunity to see the source codes in Matlab. The authors also thank Weiguo Yin from Brook Haven National Laboratory in the USA and Ruibao Tao from Fudan University in China for the opportunity to see the Fortran source codes.*

Графическая аннотация (Graphical annotation)



Верхний рисунок: пример конфигурации решетки минимального размера, в котором появляются описанные в статье барьеры.

Нижний рисунок: результат обработки этой конфигурации некорректной версией алгоритма.

Top figure: an example of a minimum lattice configuration in which the barriers described in the article appear.

Bottom figure: the result of processing this configuration by an incorrect version of the algorithm.

ВВЕДЕНИЕ

При использовании перколяционных моделей одной из важнейших задач является определение перколяционного кластера и его остова [1, 2, 3, 4]. В зависимости от моделируемой среды перколяционный кластер может проводить электрический ток, жидкость или газ. В обеспечении проводимости электрического тока, жидкости или газа участвуют не все проводящие элементы перколяционного кластера, а только те, которые входят в так называемый *остов* (backbone). Популярным вариантом перколяционных моделей являются решеточные модели. Для решеточных моделей можно определить остов как подграф перколяционного кластера, содержащий все вершины, от которых имеется как минимум два непересекающихся пути, ведущих к противоположным краям решетки.

В некоторых публикациях отмечается, что следует различать *эффективный остов* (effective backbone) и *геометрический остов* (geometric backbone) [5, 6]. Под геометрическим остовом предлагается понимать данное выше определение остова как подграфа перколяционного кластера с двумя непересекающимися путями до противоположных краев решетки или эквивалентное ему. Однако не через все элементы геометрического остова будет протекать ток из-за наличия идеально сбалансированных связей наподобие моста Уитстона. Соответственно, эффективным остовом называют только те элементы геометрического остова, через которые протекает ненулевой ток. Далее в этой статье рассматривается именно геометрический остов.

В ряде статей [7, 8, 9] рассматривается моделирование методом заливки геометрического остова в плоских квадратных решетках для задачи перколяции узлов. Хотя в указанных статьях используются разные названия – либо просто остов, либо *токонесущий остов* (current-carrying backbone) [7], либо *потоконесущий остов* (flow-carrying backbone) [8], либо *проводящий остов* (conducting backbone) [9] – фактически все три статьи рассматривают именно геометрический остов. Три упомянутые статьи рассматривают, по существу, разные версии одного и того же алгоритма.

В статьях имеются небольшие отличия в терминологии: если авторы [7] говорят об *узлах* (sites), то авторы [8] предпочитают говорить о *ячейках* (cells) на квадратной решетке. В [9] отмечается, что термины «узел» и «ячейка» здесь являются синонимами используемого в теории графов термина *вершина* (vertex), однако предпочтение в [9] отдается термину «ячейка». В плане обозначения цветами терминология авторов [7, 8, 9] очень близка: проводящие элементы называются черными узлами (ячейками), а непроводящие – белыми.

ДЕТАЛИ ОПИСАНИЯ ТРЕХ АНАЛОГИЧНЫХ АЛГОРИТМОВ

Во всех трех статьях рассматривается получение остова путем отделения от перколяционного кластера висячих частей, для которых не существует двух непересекающихся путей, ведущих к противоположным краям решетки. Еще одно небольшое техническое отличие статей [8, 9] от [7] заключается в том, что в статье [7] рассматриваются только *висячие концы* (dangling ends) и *висячие циклы* (dangling loops), а в статьях [8, 9] рассматриваются также особые висячие части, соединенные только с одной из шин, причем в [8] не используется специального названия для этих висячих частей, а в [9] эти висячие части называются висячими дугами. На самом деле термин *висячие дуги* (dangling arcs) был предложен в другой статье Иня и Тао [10], где указывалось на необходимость их отделения.

Реализация алгоритма также имеет ряд технических отличий: авторы [7] реализовывали программу на языке Fortran, авторы [8] реализовывали программу на языке Matlab, авторы [9] реализовывали программу на языке C++. Некоторые из программных кодов доступны онлайн, например, программный код, реализованный авторами [8], доступен по ссылке [11], а программный код, реализованный авторами [9], доступен по ссылке [12]. На сайте GitHub можно найти некоторые программные коды, выложенные Инь Вэйго [13]. В описании к этим кодам говорится, что они соответствуют алгоритмам статей [7] и [10]. К сожалению, данное описание неточно: в действительности коды, выложенные Инем, соответствуют статье [10]. В статье [10] рассматривается несколько другой алгоритм, ориентированный на перколяцию связей на квадратной решетке, в то время как в статье [7] рассматривается алгоритм, специфический для перколяции узлов на квадратной решетке. В выложенных программных кодах Инь реализовал для перколяции узлов применение алгоритма, описанного для перколяции связей в статье [10], основываясь на простой идее, что перколяцию узлов можно заменить на перколяцию связей, если рассматривать в перколяции связей на решетке в качестве открытых только связи, соединяющие соседние открытые узлы в соответствующем заполнении для перколяции узлов на той же решетке. Также неточным является описание файла site-percolation.zip, выложенного в [13]. Файл site-percolation.zip содержит архив с программными кодами, которые использовались Инем, чтобы сравнить реализацию предложенного им в [10] алгоритма с реализацией нахождения остова на основе алгоритма Тарьяна для нахождения двусвязных компонент графа [14]. При этом в качестве реализации на основе алгоритма Тарьяна Инем были использованы программные коды, которые написал Джонатан Гудман из Курантовского института математических наук для статьи 1988 года, опубликованной совместно с Эдвардсом и Сокалом [15].

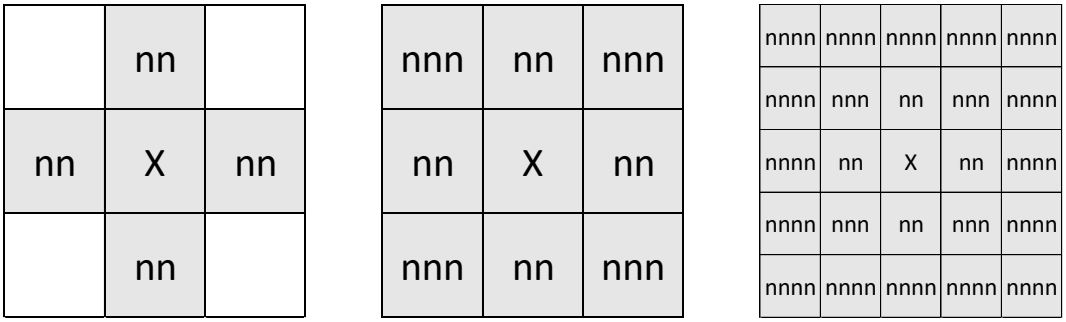
Имеются технические отличия в том, какой диапазон целых чисел используется для пометки различных типов узлов, в частности, в [7] для пометки непроводящих кластеров предлагается использовать отрицательные числа, а в [8, 9] предлагается использовать для той же цели положительные числа из диапазона, который не был задействован для других целей. Можно отметить также небольшие отличия в терминологии по поводу пометки различных типов узлов: если в [7] говорится о *метках* (labels), которыми помечаются проводящие и непроводящие узлы, то в [8, 9] метки попросту называются *цветами* (colors).

Также имеются технические отличия в том, какой способ применяется для реализации заливки: если в [8, 9] для реализации разметки кластеров заливкой используются многократные проходы по решетке до тех пор, пока меняются метки узлов, аналогично алгоритму, предложенному Штауфером для описания лесных пожаров [1, с. 5–8], то в [7] предлагается использовать для разметки кластеров более эффективный алгоритм Хошена – Копельмана [16], который, по существу, выполняет действия, аналогичные заливке, но за один проход по решетке.

Во всех трех версиях алгоритма используется понятие *сочленяющего узла* (articulation site) [7] или *сочленяющей ячейки* (articulation cell) [8, 9] перколяционного кластера – проводящего узла (ячейки), при удалении которого перколяционный кластер разрывается на две или более частей, по крайней мере, одна из которых не связана ни с одним из противоположных краев решетки (эта часть отделена и от верха, и от низа решетки). В [9] имеется небольшое отличие в определении сочленяющих ячеек: отсутствует уточнение об отсутствии связи одной из частей с краем решетки. По существу, определение сочленяющих ячеек в [9] совпадает с общим определением *точек сочленения* (articulation point) или *разделяющих вершин* (cutvertex) в теории графов [17, с. 28; 18, с. 10, 299; 19, с. 25]. Однако в реализации [12] алгоритма по статье [9] реально рассматриваются только такие точки сочленения в перколяционном кластере, при удалении которых получается часть, не соединенная ни с одним из краев. Таким образом, хотя формально в [9] определение сочленяющей

ячейки отличается от определения сочленяющего узла (сочленяющей ячейки) в [7, 8], фактически в [9] в качестве сочленяющей ячейки используется понятие, равносильное определению сочленяющего узла в [7] и сочленяющей ячейки в [8].

Во всех трех статьях [7, 8, 9] описывается один и тот же подход к анализу сочленяющих узлов (ячеек), а именно рассматриваются проводящие *nn-* и непроводящие *nnn-кластеры* (nn- и nnn-clusters), состоящие из цепочек однотипных соединенных узлов. Для определения кластеров используется понятие *окрестности* (neighborhood) узлов. В nn- и nnn-кластерах узел считается соединенным с другими узлами в пределах его nn- и nnn-окрестностей, включающих четыре или восемь соседних узлов соответственно. Аббревиатура nn происходит от английского словосочетания *nearest neighbors* (ближайшие соседи), а аббревиатура nnn происходит от словосочетания *next nearest neighbors* (следующие ближайшие соседи). Понятия ближайших соседей и следующих ближайших соседей использовались в классической работе фон Неймана по теории самовоспроизводящихся автоматов [20, с. 307; 21, с. 328], хотя аббревиатуры nn и nnn в работе фон Неймана не использовались. В [9] отмечается, что понятие nn-окрестности соответствует в теории клеточных автоматов понятию *окрестности фон Неймана* (von Neumann neighborhood), включающей 4 соседних ячейки, а понятие nnn-окрестности – *окрестности Мура* (Moore neighborhood), включающей 8 соседних ячеек [22, с. 60; 23, с. 61–62; 24, с. 38]. В [8] предлагается для анализа сочленяющих ячеек использовать также *nnnn-окрестность* (neighborhood); соответствующую аббревиатуру Тробек и Стаматович предлагают расшифровывать как *not next nearest neighbors* (не следующие ближайшие соседи). Предлагаемая в [8] nnnn-окрестность включает уже 24 соседних ячейки и соответствует *расширенной окрестности Мура* (extended Moore neighborhood) с радиусом $r = 2$ [24, с. 38]. На рисунке 1 помечены серым цветом соответствующие окрестности фон Неймана и Мура, буквой «X» помечена центральная ячейка окрестности.



а) б) в)

Рисунок 1 – Обозначения соседних ячеек и окрестности: а) окрестность фон Неймана (nn-окрестность); б) окрестность Мура (nnn-окрестность); в) расширенная окрестность Мура с радиусом $r = 2$ (nnnn-окрестность)

Также отметим, что в [7] напрямую не используется термин «окрестность», а говорится об *nn-узлах* (nn-sites), находящихся на *nn-расстоянии* (nn-distance), что соответствует nn-окрестности, и о множестве nn- и *nnn-узлов* (nnn-sites), что соответствует nnn-окрестности; при этом nnn-узлы находятся на *nnn-расстоянии* (nnn-distance). Понятие узлов с определенным числом букв «n» не тождественно понятию окрестности с таким же числом букв «n», поскольку в окрестность включаются также узлы с меньшим числом букв «n». На рисунке 1б показана nnn-окрестность и входящие в нее nn- и nnn-узлы.

В [8] в качестве синонимов понятий nn- и nnn-узлов используются соответственно понятия *nn-* и *nnn-ячеек* (nn- и nnn-cells). Также в качестве синонимов понятий nn- и nnn-узлов в [8, 9] используются соответственно термины *nn-* и *nnn-соседи* (nn- и nnn-neighbors), хотя такие термины являются не совсем аккуратными, поскольку в аббревиатуры nn- и nnn- уже входит слово *сосед* (neighbor). В некоторых источниках nn-узлы называются *первыми соседями* (first neighbours), а nnn-узлы называются *вторыми соседями* (second neighbours) [25]. Кроме того, в [8] в связи с использованием nnnn-окрестностей говорится о *nnnn-ячейках* (nnnn-cells) или *nnnn-соседях* (nnnn-neighbors). На рисунке 1в показана nnnn-окрестность и входящие в нее nn-, nnn- и nnnn-ячейки.

Отметим, что в [8] встречается смешение понятий nn-окрестности и nn-соседей, например, говорится, что «две ячейки $c_{i,j}$ и $c_{k,l}$ являются nn-соседями, если $|i-k|+|j-l| \leq 1$» («two cells $c_{i,j}$ and $c_{k,l}$ are nn-neighbors if $|i-k|+|j-l| \leq 1$»). Возможно, здесь в [8] подразумевалась nn-окрестность, иначе получается, что ячейка является сама своим nn-соседом (случай, когда $i = k$, $j = l$ удовлетворяет условию $0 \leq 1$). Также следует отметить, что в [8, с. 41] встречаются опечатки, когда nnn-кластеры называются nn-кластерами (в комментариях к шагу 1 и шагу 2 псевдокода), а в [9, с. 104–105] встречаются опечатки, когда nn-соседи называются nnn-соседями (в подписях к рисункам 12 и 13 с кодом и в тексте, где упоминаются эти рисунки).

Из текста [8] не совсем ясно, включается ли центральная ячейка, о соседях которой идет речь, в nn-, nnn- и nnnn-окрестности, поскольку везде, где в [8] упоминается слово «окрестность», речь идет только о соседях, но не о самой ячейке. В то же время в [9] явно говорится, что сама центральная ячейка в окрестность включается. На наш взгляд, естественно включать центральную ячейку в окрестность, поскольку такой подход используется и в других источниках, аккуратно систематизирующих понятия окрестностей фон Неймана и Мура [26, 27].

Для описания обработки конкретных конфигураций ячеек во всех трех статьях [7, 8, 9] используются однотипные обозначения восьми соседей центральной ячейки по сторонам света, показанные на рисунке 2. Центральная ячейка на рисунке 2 помечена буквой X, одной буквой помечены северный, западный, южный и восточный соседи (N, W, S, E соответственно от английских названий сторон света North, West, South, East), двумя буквами помечены северо-западный, юго-западный, юго-восточный и северо-восточный соседи (NW, SW, SE, NE соответственно). В статьях [8, 9] используются заглавные буквы для обозначения сторон света, а в статье [7] жирные строчные буквы (**n**,**w**,**s**,**e**).

NW	N	NE
W	X	E
SW	S	SE

Рисунок 2 – Обозначение соседних ячеек по четырем сторонам света

Висячие концы и висячие циклы присоединяются к перколяционному кластеру как раз именно через сочленяющие узлы (ячейки). Все три версии алгоритма различают два вида сочленяющих узлов, хотя в статьях используются разные термины. В статье [9] говорится о *двух видах сочленяющих ячеек*, в статье [8] говорится о *двух случаях классификации сочленяющих ячеек* (two cases in the classification of articulation cells), а в статье [7] говорится о *двух способах удаления сочленяющего узла* (two ways to remove an articulation site).

Сочленяющие узлы *первого вида* не входят ни в какой простой цикл из проводящих узлов. Обработка сочленяющих узлов первого вида во всех трех статьях [7, 8, 9] описывается практически одинаково. Эти сочленяющие узлы не могут принадлежать остову и красятся цветом непроводящего кластера вокруг. Все три описания алгоритма приравнивают к сочленяющим узлам первого вида висячие черные узлы, удаление которых не разрывает перколяционный кластер на части, но которые соединены с перколяционным кластером через один nn-узел.

Сочленяющие узлы *второго вида* входят в некоторый простой цикл из проводящих узлов. Для этих узлов сразу ответить на вопрос об их принадлежности остову нельзя, они предварительно помечаются (закрашиваются) значением (цветом) 0 (белым). На рисунке 3 показаны конфигурации, соответствующие сочленяющим узлам второго вида. На рисунке 3 метка «1» используется для проводящих узлов черного цвета, метка «*» – для узлов окружающего непроводящего кластера, метка «X» – для узлов любого цвета (проводящих или непроводящих) и метка «O» – для узлов любого цвета, отличающегося от «*» (могут быть проводящими или непроводящими). Центральный узел на рисунке 3 рассматривается как сочленяющий узел второго вида.

Следует отметить небольшие отличия в описании конфигураций узлов второго вида в [7] от [8, 9]. Для сочленяющих узлов второго вида в [8, 9] приводятся конфигурации, изображенные на рисунке 3а, 3б, 3в, а в [7] приводятся конфигурации, изображенные на рисунке 3а, 3г. Можно заметить, что конфигурация на рисунке 3б является зеркальным отражением относительно горизонтальной оси для рисунка 3а, возможно, по этой причине в [7] отдельно конфигурация рисунка 3б не приводится. Отличие рисунка 3в от рисунка 3г заключается в том, что на рисунке 3в узел NE может быть помечен тем же цветом, что и непроводящий кластер, содержащий узлы NW и SE, а на рисунке 3г узел NE обязательно отличается. Последняя особенность оговорена в тексте [7], поскольку при совпадении узлы N и E будут считаться сочленяющими ячейками первого вида, закрасятся цветом окружающего непроводящего кластера «*», и особого смысла в рассмотрении центральной ячейки как сочленяющей тогда нет. Также во всех описаниях алгоритма предполагается рассмотрение поворотов конфигураций рисунка 3 на 90, 180 и 270 градусов, хотя явно в [7] о поворотах не говорится.

X	1	O
*	1	1
X	1	*

а)

X	1	*
*	1	1
X	1	O

б)

*	1	X
1	1	1
O	1	*

в)

*	1	O
1	1	1
O	1	*

г)

Рисунок 3 – Сочленяющие ячейки второго вида

На рисунке 4 представлена обобщенная блок-схема алгоритма из статей [7, 8, 9]. Красным выделены проблемные части алгоритма, которые требуют исправления в связи с некорректной работой на некоторых конфигурациях перколяционного кластера. Синим выделен блок «отделение висячих дуг», описание которого отсутствует в публикации [7], но имеется в публикациях [8, 9], хотя в публикации [8] соответствующие части называются не висячими дугами, а *ячейками, которые nn-соединены с верхними (нижними) граничными ячейками, но не находятся в остове* (cells that are nn-connected with top (bottom) boundary cells but are not in the backbone). Аналогичный блок, называемый «идентификацией висячих дуг», имеется в алгоритме, описанном для перколяции связей в [10].

На рисунках 5–11 проиллюстрированы этапы работы приведенного на блок-схеме алгоритма для конкретного заполнения решетки. Иллюстрации сделаны на основе применения программы [12] к заполнению решетки, изображенному на рисунке 5. Рисунок 5 показывает пример начального заполнения, а рисунки 6–11 демонстрируют обработку данного заполнения решетки. Рисунок 6 иллюстрирует пометку непроводящих nnn-кластеров.

На рисунке 7 представлена пометка сочленяющих ячеек. Две ячейки на рисунке 7, в которых по сравнению с рисунком 6 метка 1 заменена на 225 (выделено темно-зеленым цветом), являются соответственно сочленяющей ячейкой первого вида (нижняя ячейка) и висячей ячейкой (верхняя ячейка). Одиннадцать ячеек, помеченных нулями на рисунке 7, являются сочленяющими ячейками второго вида.

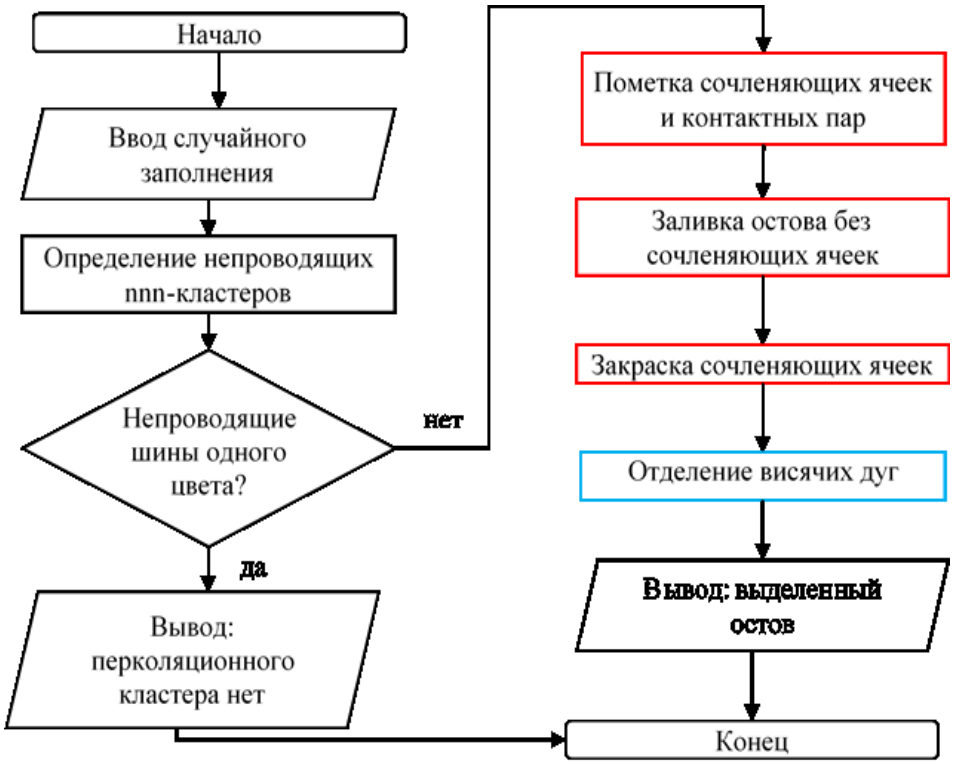


Рисунок 4 – Обобщенная блок-схема алгоритма

0	1	1	1	1	1	1	1	1	1	1	1	1	0
0	1	0	0	1	0	0	0	0	0	0	1	0	0
0	1	0	1	1	0	1	1	0	1	1	1	0	0
0	1	1	1	0	0	1	1	1	1	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	1	1	0
0	0	0	0	0	0	1	1	1	1	0	1	1	0
0	0	0	0	0	0	1	1	0	1	1	1	0	0
0	1	0	1	1	0	0	0	0	1	1	1	0	0
0	1	0	1	1	1	1	1	0	1	0	1	1	0
0	1	1	0	1	0	0	1	0	1	0	1	1	0
0	1	1	1	1	1	0	1	1	1	1	0	0	0
0	0	1	0	1	1	0	1	0	1	1	0	0	0
0	0	1	0	0	0	0	1	0	0	0	0	0	0
0	1	1	1	1	1	1	1	1	1	1	1	1	0

Рисунок 5 – Начальное состояние решетки после ввода заполнения проводящими и непроводящими узлами

После пометки сочленяющих ячеек второго вида остов может оказаться разбит на несколько проводящих nn-кластеров, поскольку сочленяющие ячейки второго вида на этом этапе исключаются из проводящих. Проводящие nn-кластеры соединяются между собой через контактные пары, то есть пары проводящих узлов, которые являются друг для друга nnn-узлами (имеют общий угол), а для сочленяющих ячеек nn-узлами. На рисунке 8 ячейки, входящие в контактные пары, соединены коричневыми отрезками. Следует отметить, что имеются технические отличия в пометке контактных пар в статьях [8] и [9]. В [8] предлагается пометка контактных пар особыми цветами ячеек, которые зависят от расположения контактной пары и отличаются от других цветов, при этом ячейки, входящие в одну контактную пару, помечаются одним цветом. В [9] отмечается, что такой подход к пометке контактных пар может приводить к недоразумениям, поскольку одна и та же ячейка может входить в более чем одну контактную пару, и предлагается хранить информацию о контактных парах в отдельной структуре данных. На рисунке 8 показаны три ячейки, входящие в две контактные пары одновременно (11-я строка, 4-й столбец; 10-я строка, 5-й столбец; 5-я строка, 8-й столбец). В [7] говорится о пометке контактных пар, но технические детали хранения информации о контактных парах не уточняются и, ввиду недоступности исходного кода программы, неизвестны.

225	1	1	1	1	1	1	1	1	1	1	1	1	238
225	1	51	51	1	225	225	225	225	225	225	1	238	238
225	1	51	1	1	225	1	1	225	1	1	1	238	238
225	1	1	1	225	225	1	1	1	1	238	238	238	238
225	225	225	225	225	225	225	1	238	238	238	1	1	238
225	225	225	225	225	225	1	1	1	1	238	1	1	238
225	225	225	225	225	225	1	1	225	1	1	1	238	238
225	1	225	1	1	225	225	225	225	1	1	1	238	238
225	1	225	1	1	1	1	1	225	1	238	1	1	238
225	1	1	225	1	215	215	1	225	1	238	1	1	238
225	1	1	1	1	1	215	1	1	1	1	238	238	238
225	225	1	215	1	1	215	1	238	1	1	238	238	238
225	225	1	215	215	215	215	1	238	238	238	238	238	238
225	1	1	1	1	1	1	1	1	1	1	1	1	238

Рисунок 6 – Состояние решетки после определения непроводящих nnn-кластеров

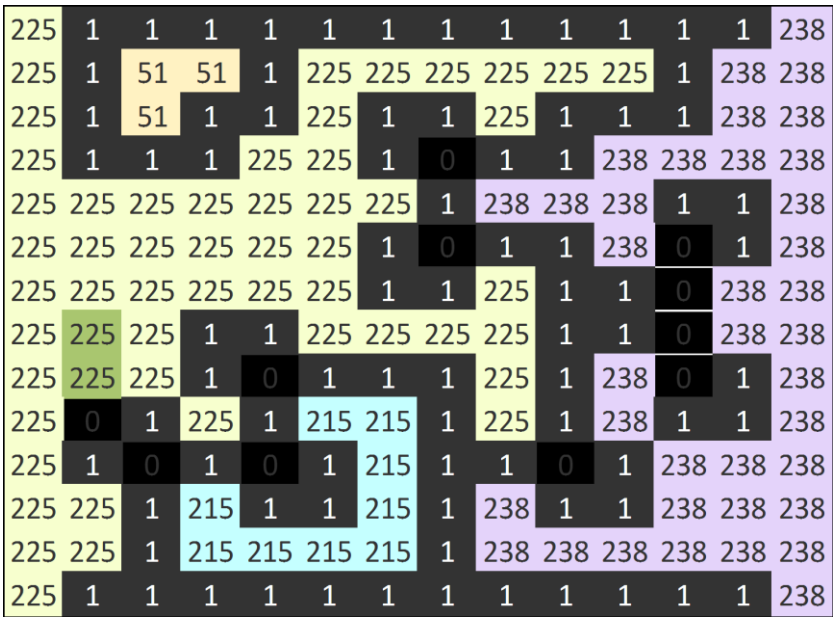


Рисунок 7 – Состояние решетки после пометки сочленяющих ячеек

На следующем этапе алгоритма (заливка остова без сочленяющих ячеек) контактные пары используются для соединения частей остова. На рисунке 9 можно увидеть, какие контактные пары были использованы для соединения nn-кластеров при заливке остова, которая показана зеленым цветом «–2». На рисунке 10 показана закраска «–2» входящих в остов сочленяющих ячеек на следующем этапе.

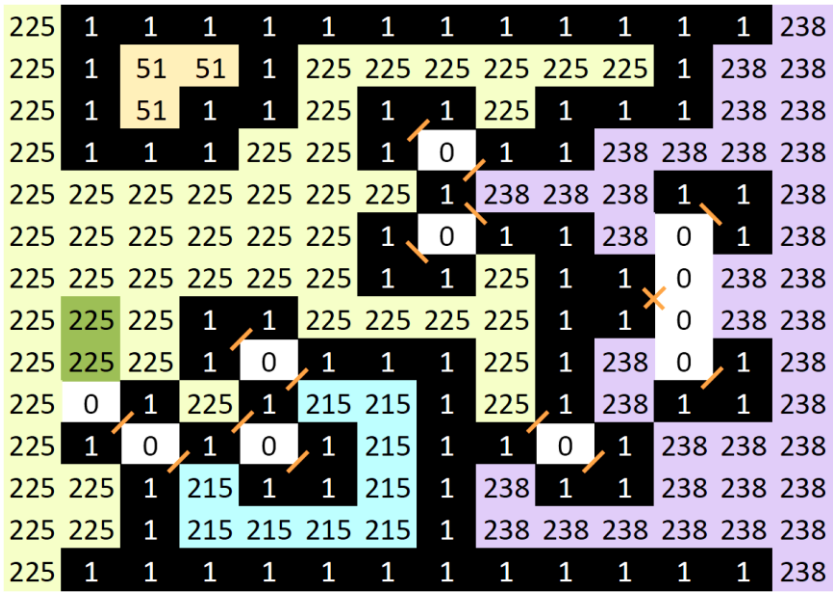


Рисунок 8 – Состояние решетки после пометки сочленяющих ячеек с пометкой контактных пар

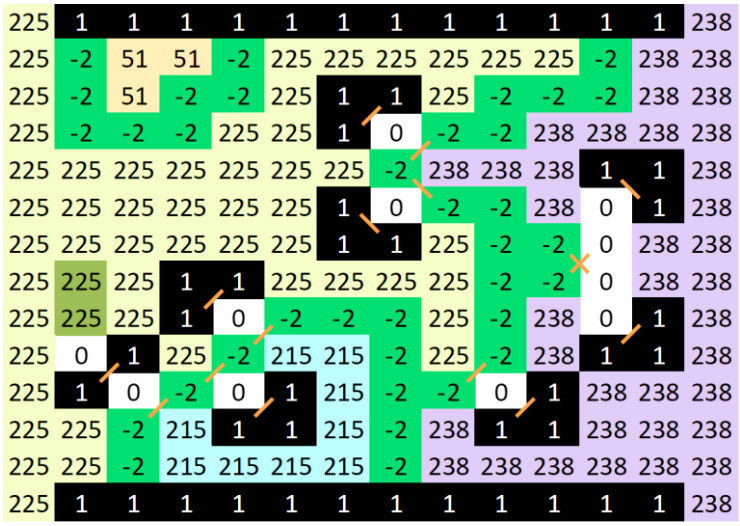


Рисунок 9 – Заливка остова с протеканием через контактные пары

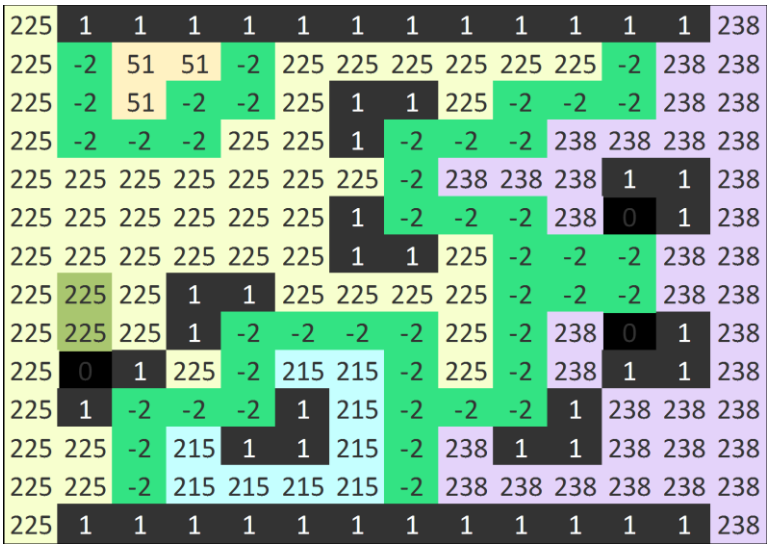


Рисунок 10 – Закраска, входящих в остов сочленяющих ячеек

На рисунке 11 показан результат описанного только в [8] и [9] последнего этапа алгоритма: отделение висячих дуг. Получившийся окончательный остов помечен красным «–3».

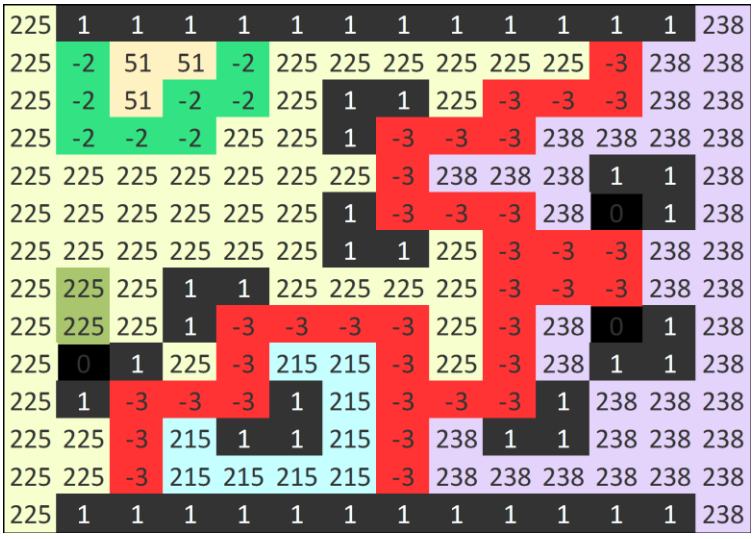


Рисунок 11 – Отделение висячих дуг

ПРОБЛЕМЫ В ОПИСАННЫХ РАНЕЕ АЛГОРИТМАХ

Тробек и Стаматович [8], а также Гордеев с Овчаренко и Сизовой [9], описывая алгоритм, не предусмотрели, что могут образовываться барьеры из сочленяющих ячеек. Если сочленяющая ячейка входит в контактную пару, то заливка через нее не проходит. Использование контактных пар не позволяет преодолеть сложные барьеры из сочленяющих ячеек.

Проблема заливки с контактными парами иллюстрируются далее на основе показанного на рисунке 12 примера конфигурации. На рисунке 13 показан результат пометки сочленяющих ячеек и контактных пар конфигурации рисунка 12 согласно реализации алгоритма, описанной в [9]. В показанной конфигурации возникают два барьера из четырех сочленяющих ячеек (четыре рядом стоящих нуля), которые разрезают перколяционный кластер. При этом две центральных сочленяющих ячейки барьеров одновременно входят в контактные пары. На рисунках 14 и 15 показано, что заливка от верхней и нижней шины доходит до этих барьеров, но не перетекает через них. В результате при заливке не возникает целого кластера, соединяющего верхнюю и нижнюю шины, а залитые значением «–2» части перколяционного кластера отбрасываются на этапе отделения висячих дуг. Реализация алгоритма, описанная в [8], работает похожим образом, отличие только в том, что в этой реализации центральные ячейки барьеров даже не помечаются как входящие контактные пары.

0	1	1	1	1	1	1	1	1	1	0
0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	1	0	1	1	0	0	0
0	0	1	1	1	1	1	1	0	0	0
0	0	1	1	0	1	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	1	0	1	1	0	0
0	0	0	1	1	1	1	1	1	0	0
0	0	0	1	1	0	1	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0
0	1	1	1	1	1	1	1	1	1	0

Рисунок 12 – Конфигурация, в которой возникают барьеры из сочленяющих ячеек

В описании алгоритма у Иня и Тао [7, с. 85] имеется странное место, которое может приводить к тому, что пометка сочленяющих узлов второго вида и/или соответствующих контактных пар будет зависеть от порядка обработки узлов. В описании порядка обработки сочленяющего узла, соответствующего рисунку 3а, говорится следующее: «Тогда *a* удаляется «временно» путем установки нуля в качестве метки *a*. Кроме того, **e** и **n** помечаются как контактная пара. Однако если метка **n** или метка **e** равна нулю, то мы ничего не делаем для этого случая». (Then *a* is removed “temporarily” by setting zero as the label of *a*. Besides, **e** and **n** are tagged as a contact couple. However, if the label of **n** or the label of **e** is zero, then we do nothing for this case.) Здесь не совсем ясна фраза «мы ничего не делаем для этого случая» (we do nothing for this case). Означает ли фраза «ничего не делаем» только, что не делается пометка контактной пары, или же не делается также пометка сочленяющего узла *a* нулем? Авторы [8] и [9] реализовали сначала пометку сочленяющих ячеек, а затем пометку контактных пар, предположив, что фраза «ничего не делаем» у Иня и Тао относится только к пометке контактных пар, а сочленяющие ячейки надо помечать в любом случае. В результате и в [8], и в [9] возникают непреодолимые барьеры из сочленяющих ячеек, показанные на рисунках 13–15.

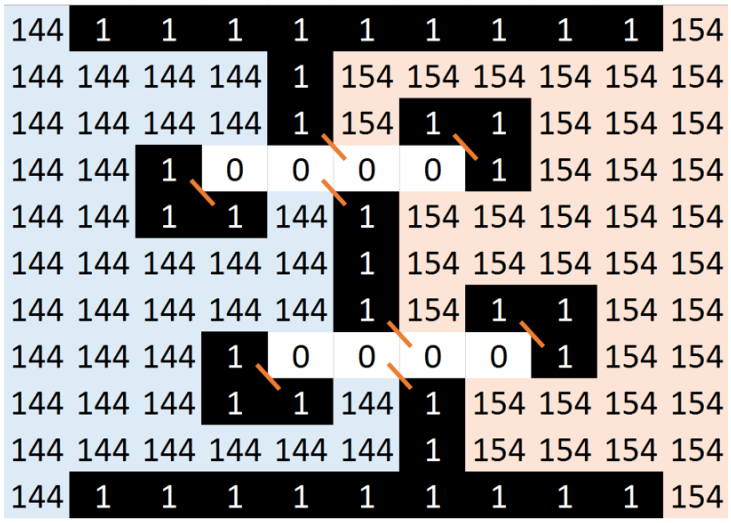


Рисунок 13 – Барьеры из сочленяющих ячеек

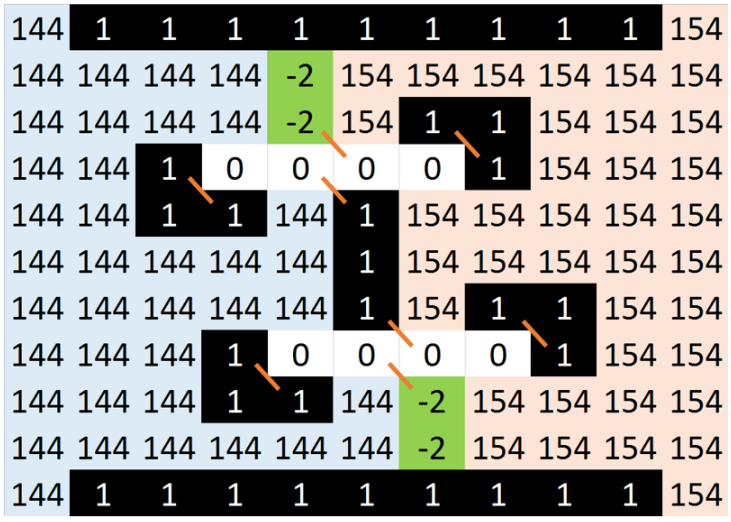


Рисунок 14 – Заливка не преодолевает барьеры

Если же понимать, что фраза «ничего не делаем» относится также к пометке сочленяющих ячеек, то это означает, что Инь и Тао предполагали здесь одновременную пометку сочленяющих ячеек и контактных пар, без разбиения на два подэтапа, как в [8] и [9]. В таком случае одна из двух центральных ячеек в каждом из двух барьеров не будет помечена как сочленяющая. Какая именно из двух центральных сочленяющих ячеек не будет помечена, зависит от порядка обработки ячеек в строке: при обработке слева направо не будет помечена правая сочленяющая ячейка, а при обработке справа налево не будет помечена левая сочленяющая ячейка. Независимо от того, какая из двух центральных сочленяющих ячеек не будет помечена, другая центральная сочленяющая ячейка даст контактную пару, по которой протечет заливка (произойдет соединение частей перколяционного кластера). Возможно, что Инь и Тао заметили проблему с образованием барьеров из сочленяющих ячеек и решили при помощи такого специфического подхода решить данную проблему.

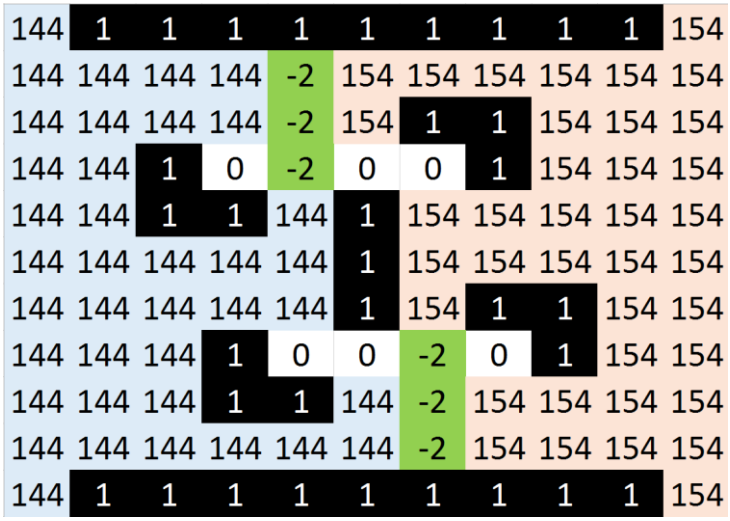


Рисунок 15 – Результат присоединения сочленяющих ячеек к остову

На первый взгляд может показаться, что, решив подобным образом проблему с барьерами, Инь и Тао получили корректный алгоритм выделения остова. Однако, если использовать такое понимание алгоритма, что сочленяющие ячейки должны помечаться не всегда, то, к сожалению, это приводит не только к преодолению барьеров, но также к присоединению некоторых висячих циклов.

На рисунке 16 показан пример конфигурации заполнения решетки, для которой возможно присоединение висячих циклов, если понимать, что в алгоритме Иня и Тао помечаются не все сочленяющие ячейки. На рисунке 17 помечены все сочленяющие ячейки и соответствующие контактные пары в соответствии с версией алгоритма, которая реализована в [12].

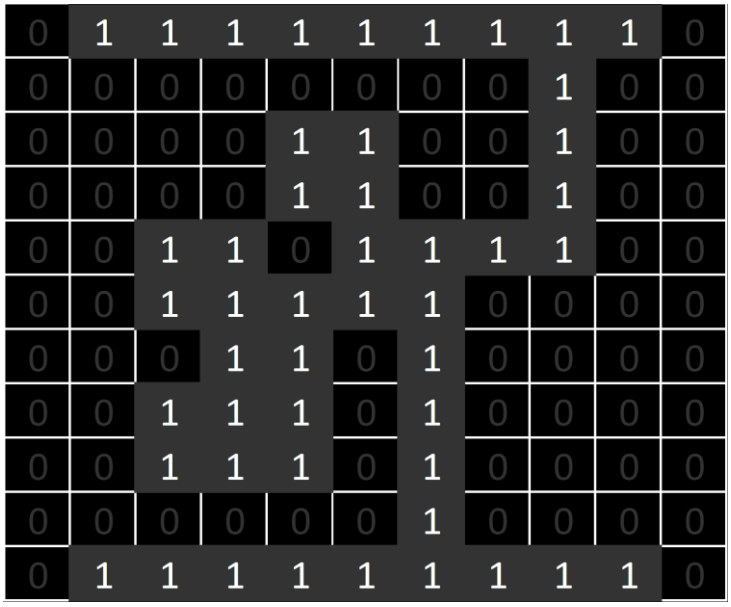


Рисунок 16 – Пример конфигурации, вызывающей проблемы, если помечать не все сочленяющие ячейки

Однако, если упомянутая выше фраза «ничего не делаем» означает, что не надо помечать сочленяющие ячейки, когда в соответствующей контактной паре есть ячейка, помеченная ранее нулем, то, при обработке ячеек по строкам слева направо и перебирая строки сверху вниз, две сочленяющие ячейки будут не помечены. В результате получится конфигурация, показанная на рисунке 18. При этом на рисунке 18 контактные пары, в которых есть ноль и из-за которых не были помечены нулем сочленяющие ячейки, показаны красными отрезками. В результате того, что две сочленяющие ячейки не помечены нулями, к остову присоединится большой висячий цикл, находящийся на рисунке 18 внизу слева. Размеры подобных висячих циклов могут быть сравнимы с размерами решетки, а поскольку подобные висячие циклы могут встречаться в разных ориентациях, то изменение порядка обработки ячеек, например, на порядок справа налево может привести к присоединению подобных висячих циклов в другой ориентации.

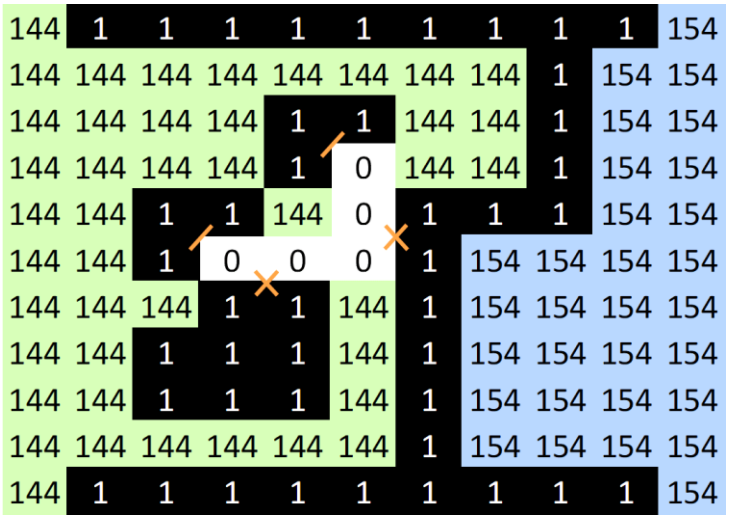


Рисунок 17 – Полная пометка сочленяющих ячеек и контактных пар

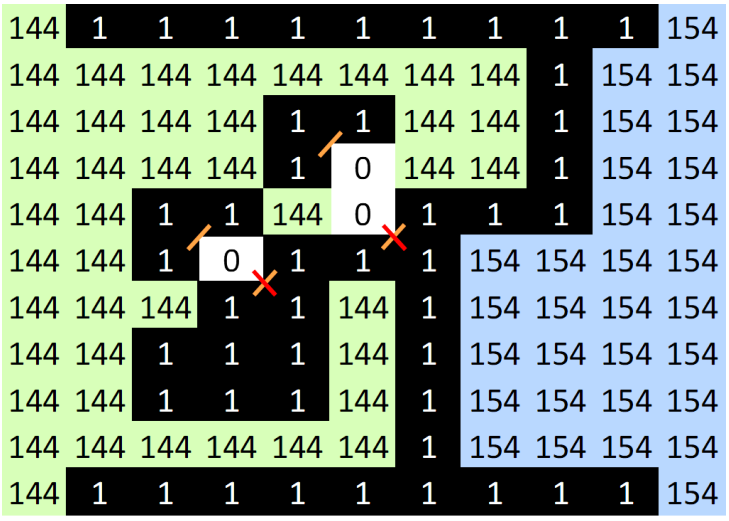


Рисунок 18 – Две сочленяющие ячейки остались непомеченными

ИСПРАВЛЕННАЯ И УПРОЩЕННАЯ ВЕРСИЯ АЛГОРИТМА

В результате анализа отмеченных недостатков версий алгоритма, описанных в [7, 8, 9], нами было предложено исправление алгоритма, которое также привело к его упрощению. На рисунке 19 представлена блок-схема упрощенного алгоритма. Предложенные изменения упрощают алгоритм и одновременно решают проблему с барьерами из сочленяющих ячеек.

Предусмотреть все возможные конфигурации барьеров из сочленяющих ячеек затруднительно. Поэтому вместо использования контактных пар предлагается использовать пометку ориентации сочленяющих ячеек второго вида. Для удобства предлагается различать сочленяющие ячейки, включенные в остов, и сочленяющие ячейки, не включенные в остов, разными пометками.

Сочленяющие ячейки могут иметь две ориентации, показывающие способ протекания через ячейку:

– протекание с юго-запада на северо-восток («–4», в остове «–6»);
– протекание с юго-востока на северо-запад («–5», в остове «–7»).

На рисунке 20 показана пометка сочленяющих ячеек в улучшенном алгоритме для заполнения решетки, показанного на рисунке 5. Заливка непроводящих кластеров в улучшенном алгоритме не отличается, результат совпадает с рисунком 6. Соответственно, рисунок 20 является аналогом рисунка 8, но на рисунке 20 отличается пометка сочленяющих ячеек второго вида. Если на рисунке 8 все сочленяющие ячейки второго вида помечены значением «0», то на рисунке 20 используется отдельное значение для пометки сочленяющих ячеек, ориентированных с юго-запада на северо-восток, «–4», и отдельное значение для пометки сочленяющих ячеек, ориентированных с юго-востока на северо-запад, «–5». Для наглядности сопоставления рисунков 8 и 20 сочленяющие ячейки, поме-

ченные этими значениями («–4» и «–5»), по-прежнему выделены белым цветом. Также для наглядности внутри сочленяющих ячеек при помощи коротких синих отрезков показана их ориентация. При этом синие отрезки касаются границ ячеек, между которыми возможно протекание (входивших ранее в контактную пару), но сами контактные пары не помечаются.

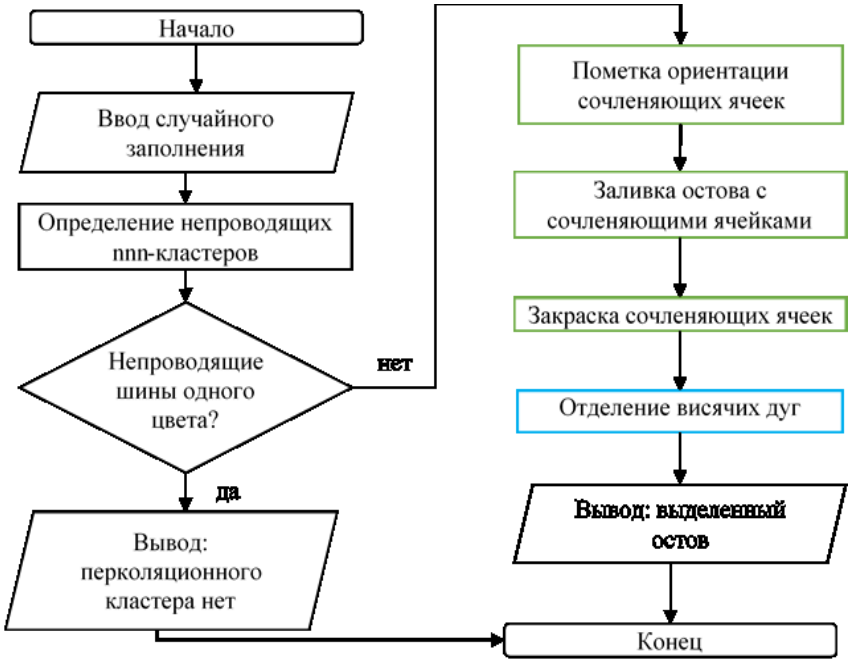


Рисунок 19 – Блок-схема упрощенного алгоритма

Затем в улучшенном алгоритме делается заливка остова с сочленяющими ячейками второго вида, входящими в остов, показанная на рисунке 21. При этом делается предварительное включение в остов сочленяющих ячеек второго вида путем пометки специальными значениями сочленяющих ячеек, ориентированных с юго-запада на северо-восток «–6» и ориентированных с юго-востока на северо-запад «–7». Рисунок 21 можно считать аналогом рисунка 9, но на рисунке 21 предварительно вошедшие в остов сочленяющие ячейки помечены светло-зеленым цветом («–6» и «–7»). Рисунок 22 является аналогом рисунка 10. Здесь сочленяющие ячейки второго вида, вошедшие в остов, помечены значением «–2» так же, как и остальные ячейки остова. Предварительная пометка этих сочленяющих ячеек специальными значениями («–6» и «–7») требовалась, чтобы сохранять информацию об ориентации этих ячеек для алгоритма заливки.

Этап отделения висячих дуг не отличается от показанного на рисунке 11, поэтому дублирующий рисунок для улучшенного алгоритма не приводится.

225	1	1	1	1	1	1	1	1	1	1	1	238
225	1	51	51	1	225	225	225	225	225	225	1	238 238
225	1	51	1	1	225	1	1	225	1	1	1	238 238
225	1	1	1	225	225	1	-4	1	1	238	238	238 238
225	225	225	225	225	225	225	1	238	238	238	1	1 238
225	225	225	225	225	225	1	-5	1	1	238	-5	1 238
225	225	225	225	225	225	1	1	225	1	1	-5	238 238
225	225	225	1	1	225	225	225	225	1	1	-4	238 238
225	225	225	1	-4	1	1	1	225	1	238	-4	1 238
225	-4	1	225	1	215	215	1	225	1	238	1	1 238
225	1	-4	1	-4	1	215	1	1	-4	1	238	238 238
225	225	1	215	1	1	215	1	238	1	1	238	238 238
225	225	1	215	215	215	215	1	238	238	238	238	238 238
225	1	1	1	1	1	1	1	1	1	1	1	1 238

Рисунок 20 – Состояние решетки после пометки ориентации сочленяющих ячеек

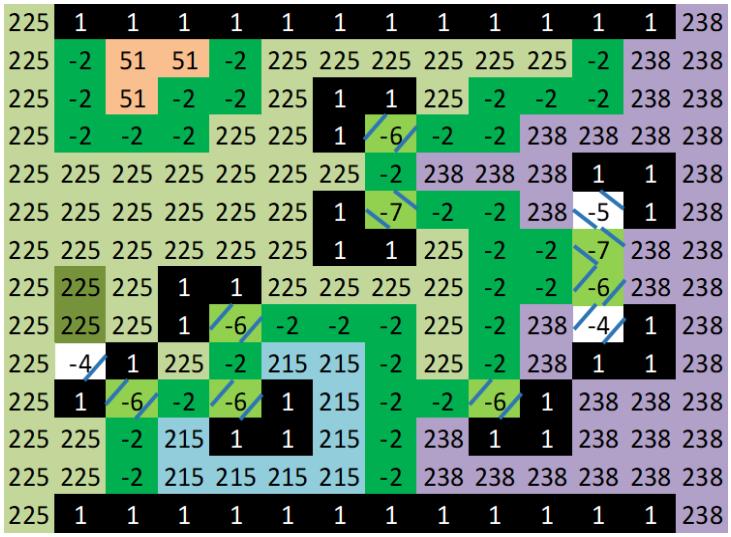


Рисунок 21 – Заливка остова с предварительным включением сочленяющих ячеек

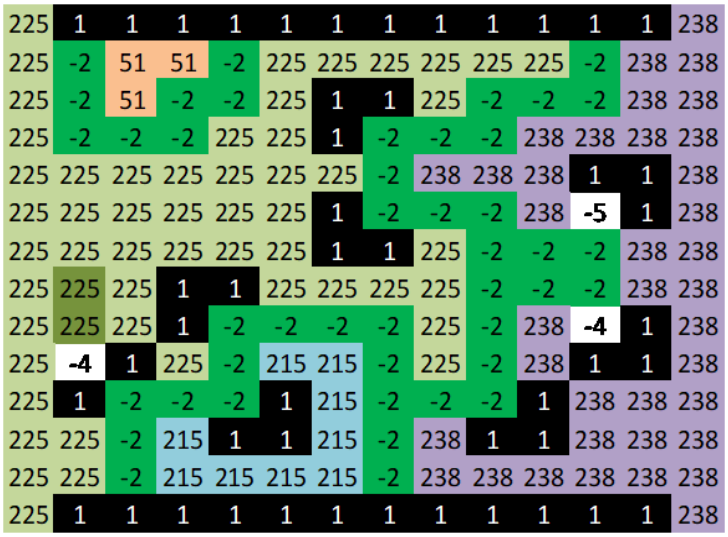


Рисунок 22 – Закраска входящих в остов сочленяющих ячеек

На рисунках 20–22 показано сравнение работы улучшенного алгоритма с прежней версией на заполнении решетки, в котором не возникало проблемы. Однако улучшенная версия алгоритма успешно справляется и с барьерами из сочленяющих ячеек, поскольку в улучшенном алгоритме протекание может происходить как между обычными и сочленяющими ячейками, так и из одной сочленяющей ячейки в другую, соседнюю с ней сочленяющую ячейку, с учетом ориентации сочленяющих ячеек.

КОРРЕКТНАЯ ОБРАБОТКА БАРЬЕРОВ ИЗ СОЧЛЕНЯЮЩИХ ЯЧЕЕК

Обработка проблемного заполнения, показанного на рисунке 12 улучшенным алгоритмом, продемонстрирована на рисунках 23–25. Рисунок 23 является аналогом рисунка 13, в нем обнаруживаются точно такие же барьеры из сочленяющих ячеек, но теперь эти сочленяющие ячейки барьеров помечены с учетом ориентации значением «–5». Синие отрезки внутри ориентированных сочленяющих ячеек показывают, между какими краями сочленяющей ячейки, может происходить протекание.

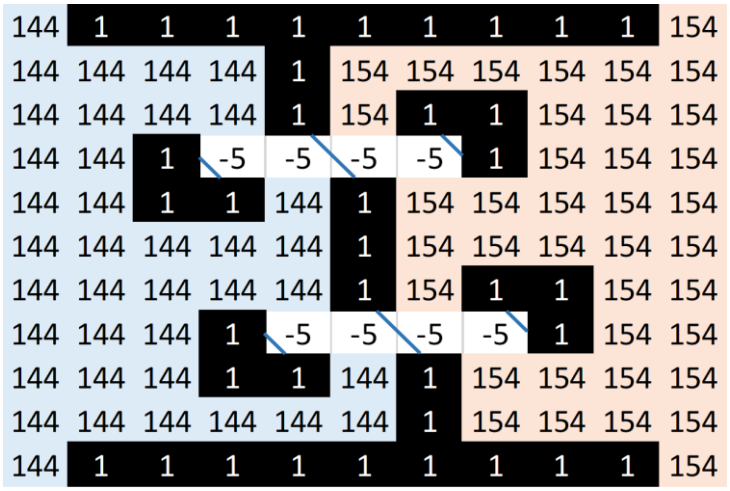


Рисунок 23 – Барьеры из сочленяющих ячеек

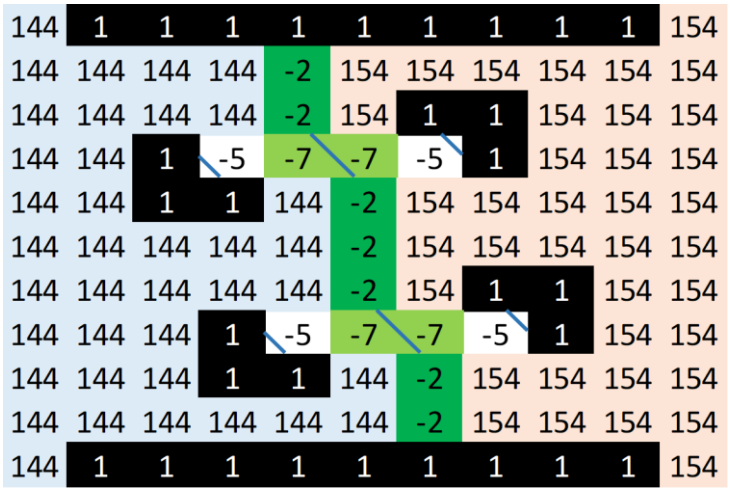


Рисунок 24 – Заливка преодолевает барьеры

Рисунок 24 является аналогом рисунка 14, но теперь заливка преодолевает барьеры через ориентированные сочленяющие ячейки. Рисунок 25 является аналогом рисунка 15, но теперь все сочленяющие ячейки, через которые прошла заливка, корректно присоединяются к остову. Висячие дуги в примере на рисунка 25 отсутствуют, поэтому следующий этап отделения висячих дуг не показан.

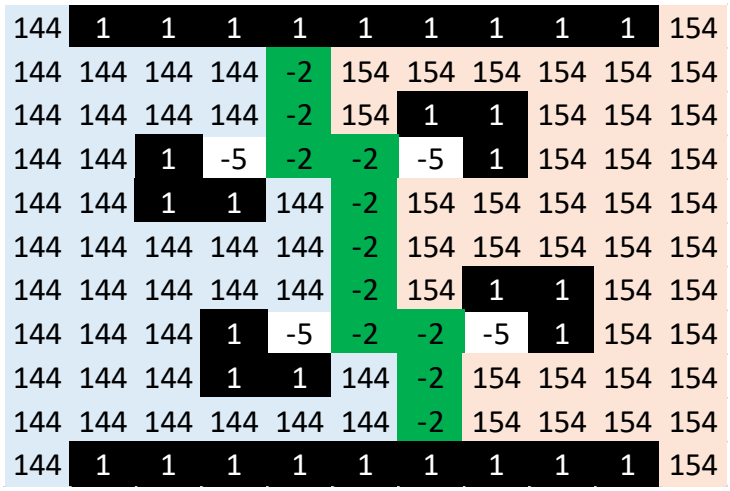


Рисунок 25 – Результат присоединения сочленяющих ячеек к остову

ЭКСПЕРИМЕНТАЛЬНОЕ СРАВНЕНИЕ ЗАЛИВКИ С ДРУГИМИ АЛГОРИТМАМИ

Было проделано экспериментальное сравнение по времени выполнения и доле отличий в остове на случайных решетках для двух алгоритмов заливки (с ошибкой по [12] и исправленная версия, описанная в данной статье [28]) с рассматривавшимися в публикации [29] реализацией алгоритма Грассбергера с четырьмя поворотами и реализацией алгоритма Ахунжанова. Аналогично статье [29] брались решетки с размерами $L = 5, 7, 10, 14, 20, 28, \ldots, 1280, 1792, 2560, 3584, 5120$. Для решетки каждого размера рассматривалось по 200 случайных реализаций. Также аналогично [29] при тестировании всех алгоритмов рассматривались случайные смачивания снизу решетки, сгенерированные с помощью программы PureGrasberger.cpp [30].

В результате тестирования было установлено, что остовы, получаемые с помощью исправленной версии алгоритма заливки [28], во всех реализациях решеток для всех размеров решеток совпадают с корректными остовами, получаемыми алгоритмом Ахунжанова. В то же время при тестировании алгоритма заливки [12], содержащего описанную в данной статье ошибку, были получены существенные отличия в меньшую сторону.

На рисунке 26 показана средняя доля отличий остовов, получаемых алгоритмом заливки с ошибкой, от остовов, получаемых алгоритмом Ахунжанова или исправленным алгоритмом заливки, на решетках с размерами от $L = 640$ до $L = 5120$. Поскольку все отклонения имели место в меньшую сторону от правильных остовов, то на графике показаны отрицательные значения от $-2{,}2$ до $-36{,}4$ %. Погрешность средней доли отличий показана на графике вертикальными штрихами.

На решетках размеров от $L = 5$ до $L = 448$ в подавляющем большинстве случаев остов определялся правильно, и доля отличия остова не превышала 1,3 % по абсолютной величине. С увеличением размера решетки L все большее количество конфигураций заполнения решеток имели отличия остова в меньшую сторону. Так, если для решеток размером $L = 640$ 16 % конфигураций заполнения решеток имели отличия остова в меньшую сторону, то при $L = 5120$ уже на всех конфигурациях остов отличался в меньшую сторону. Следует отметить, что в большинстве конфигураций имеет место лишь частичное отрезание одной из ветвей остова, когда существуют параллельные ветви остова, и на одной из параллельных ветвей остова встречаются барьеры из сочленяющих ячеек, а на другой ветви барьеров нет. Но в ряде случаев происходило и полное зануление остова, когда барьеры из сочленяющих ячеек разрывали единственную ветвь остова, как описано выше. При этом, если для решеток размером $L = 640$ полное зануление остова происходило только в 2 % конфигураций, то при $L = 5120$ полное зануление остова имело место уже в 33 % случаев при наличии перколяции. Доля конфигураций, на которых имело место полное зануление остова, существенно влияло на среднюю долю отклонения. Так, например, при стороне решетки $L = 1792$ было обнаружено 14 % конфигураций с полным занулением остова, а при $L = 2560$ только 9 % конфигураций с полным занулением остова. По этой причине на графике (рис. 26) доля отличий при $L = 1792$ оказалась чуть больше по модулю, чем при $L = 2560$, хотя в целом наблюдается тенденция к более сильным отличиям с увеличением размера решетки.

Для четырех версий программы был произведен замер времени, затрачиваемого в среднем на обработку одной решетки со стороной L узлов. Усреднение времени работы проводилось по 200 различным случайным заполнениям решетки. Однако среднее время работы ошибочной и исправленной версии алгоритма заливки практически на всех размерах решеток оказывалось одинаковым. Поэтому на рисунке 27 показаны графики зависимости среднего времени обработки решетки для трех версий программы: алгоритм Грассбергера с четырьмя ориентациями, алгоритм Ахунжанова и исправленный алгоритм заливки. Стандартная погрешность для среднего времени обработки решетки не превышает размеров маркеров. На графиках показаны размеры решеток начиная с $L = 20$, поскольку на меньших размерах решеток от $L = 5$ до $L = 14$ погрешность в замерах времени оказывалась слишком большой.

Для тестирования всех программ использовался компьютер со следующими характеристиками:

- модель процессора – Intel (R) Core (TM) i3-10110U CPU;
- частота – 2,10 GHz;
- количество ядер – 2;
- количество логических процессоров – 4;
- количество оперативной памяти – 8 Гб;
- операционная система Windows 10.

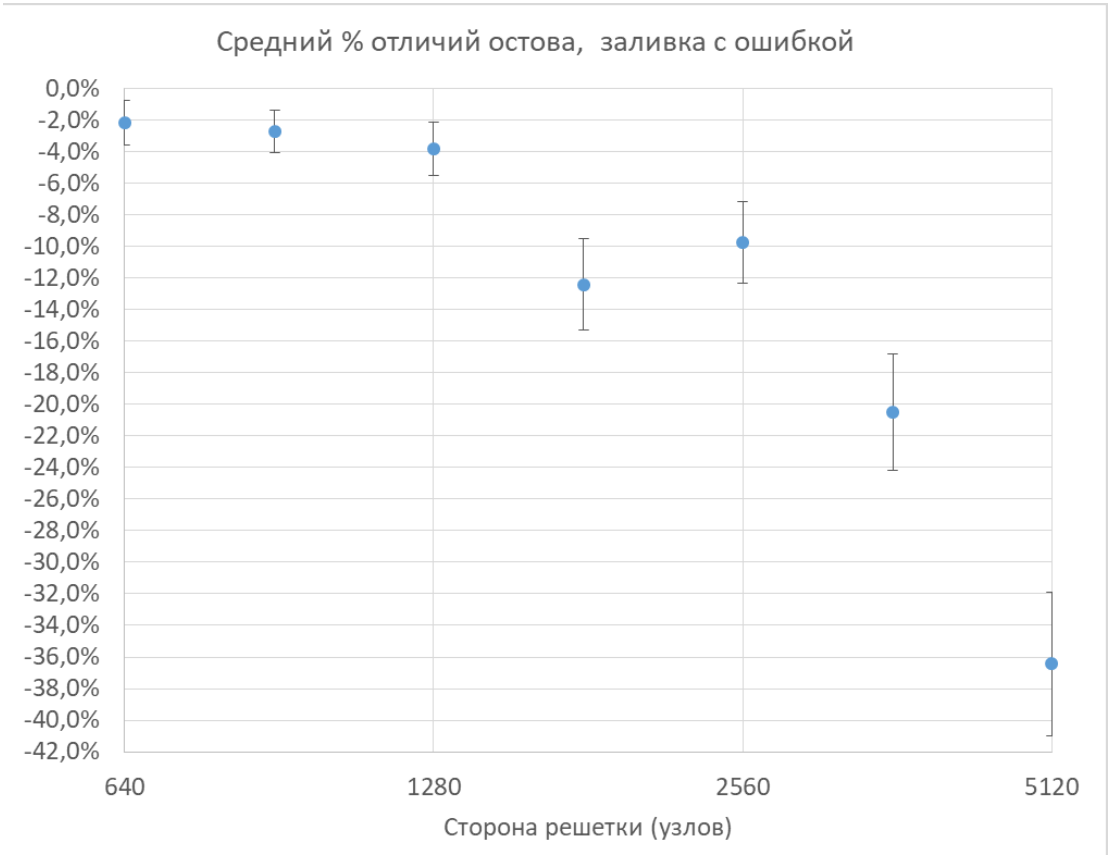


Рисунок 26 – Средний процент отличий остова, полученного алгоритмом заливки с ошибкой, от остова, полученного алгоритмом Ахунжанова для разных размеров решеток

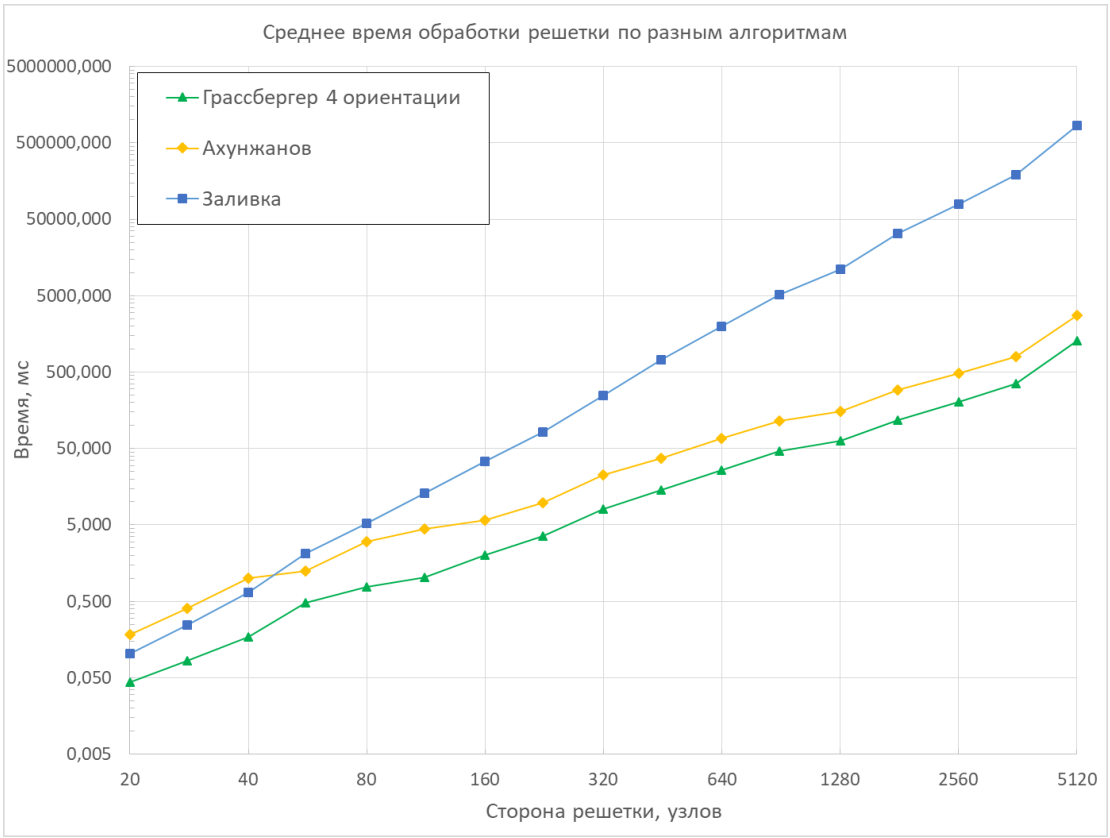


Рисунок 27 – График среднего времени обработки решетки для разных алгоритмов в зависимости от стороны решетки

ЗАКЛЮЧЕНИЕ

Предлагавшийся ранее в [8, 9] алгоритм заливки содержит существенную ошибку, в результате которой на больших решетках средняя доля отличий от правильного остова превышает 35 %, что вряд ли можно считать приемлемым даже для грубых оценок. Еще менее приемлемым следует считать полное зануление остова, поэтому от использования для определения остова алгоритма, описанного в [8, 9], следует отказаться. Исправленный алгоритм заливки, предложенный в данной статье, дает корректный остов, совпадающий с остовом, получаемым алгоритмом Ахунжанова. Однако по времени работы исправленный алгоритм заливки оказывается быстрее алгоритма Ахунжанова только на малых решетках со стороной до $L \leq 40$, а на больших решетках быстрее оказывается алгоритм Ахунжанова за счет лучшей асимптотической сложности, рассматривавшийся в [29]. Худшая асимптотическая сложность алгоритма заливки, отмечавшаяся в [9], связана с полным просмотром решетки для реализации очередного шага заливки. Если отказаться от многократных просмотров решетки, то, возможно, асимптотическая сложность алгоритма заливки может быть улучшена, о чем упоминается в [7], однако этот вопрос требует отдельного рассмотрения.

Список источников

1. Stauffer, D. Introduction to Percolation Theory / D. Stauffer and A. Aharony. – 2nd rev. ed. – London : Taylor & Francis, 2003. – 180 p.

2. Grimmet G. Percolation / Geoffrey Grimmett. – 2nd ed. – Berlin ; Heidelberg ; New York ; Barcelona ; Hong Kong ; London ; Milan ; Paris ; Singapore ; Tokyo : Springer, 1999. (Grundlehren der mathematischen Wissenschaften; 321).

3. Тарасевич, Ю. Ю. Перколяция: Теория, приложения, алгоритмы : учебное пособие / Ю. Ю. Тарасевич. – Изд. 2-е. – Москва : Книжный дом «ЛИБРОКОМ», 2012. – 112 с.

4. Sahimi, M. Percolation Phase Transition / M. Sahimi // Complex Media and Percolation Theory. – New York : Springer Nature, 2021. – P. 1–9.

5. Roux, S. A new algorithm to extract the backbone in a random resistor network / S. Roux, A. Hansen // Journal of Physics A: Mathematical and General. – 1987. – Vol. 20. – P. L1281–L1285.

6. Tarasevich, Yu. Yu. Identification of current-carrying part of a random resistor network: electrical approaches vs. graph theory algorithms / Yu. Yu. Tarasevich, A. S. Burmistrov, V. A. Goltseva, I. I. Gordeev, V. I. Serbin, A. A. Sizova, I. V. Vodolazskaya and D. A. Zholobov // Journal of Physics: Conference Series. – 2018. – Vol. 955. – P. 012021. – DOI: 10.1088/1742-6596/955/1/012021.

7. Yin, W.-G. Algorithm for finding two-dimensional site percolation backbones / W.-G. Yin, R. Tao // Physica B: Condensed Matter. – 2000. – Vol. 279. – P. 84–86.

8. Trobec, R. Analysis and classification of flow-carrying backbones in two-dimensional lattices / R. Trobec, B. Stamatovic // Advances in Engineering Software. – 2017. – Vol. 103. –P. 38–45.

9. Гордеев, И. И. Нахождение проводящего остова в двумерной решетке методом заливки / И. И. Гордеев, С. С. Овчаренко, А. А. Сизова // Прикаспийский журнал: управление и высокие технологии. – 2020. – № 1 (49). – С. 94–111.

10. Yin, W.-G. Rapid algorithm for identifying backbones in the two-dimensional percolation model / W.-G. Yin, R. Tao // International Journal of Modern Physics C. – 2003. – Vol. 14. – P. 1427–1437. – DOI: 10.1142/S0129183103005509.

11. Trobec, R. Index of /~roman/files/tmp/backbone / R. Trobec // Jozef Stefan Institute. – URL: https://e6.ijs.si/~roman/files/tmp/backbone/ (дата обращения: 09.12.2022).

12. G2ii2g. G2ii2g/backbone-flooding: Program, tests and results for flooding of backbone on square lattice // GitHub. – URL: https://github.com/G2ii2g/backbone-flooding (дата обращения: 09.12.2022).

13. yinweiguo. yinweiguo/FindingBackbone: A program to find the current-carrying backbones in the two-dimensional percolation model // GitHub. – URL: https://github.com/yinweiguo/FindingBackbone (дата обращения: 09.12.2022).

14. Tarjan, R. Depth-first search and linear graph algorithms / R. Tarjan // SIAM Journal on Computing. – 1972. – Vol. 1. – P. 146–160.

15. Edwards, R. Multigrid Method for the Random-Resistor Problem / R. Edwards, G. J. Goodman and A. D. Sokal // Physical Review Letters. – 1988. – Vol. 61. – P. 1333–1335.

16. Hoshen, J. Percolation and cluster distribution. I. Cluster multiple labeling technique and critical concentration algorithm / J. Hoshen and R. Kopelman // Physical Review B. – 1976. – Vol. 14. – P. 3438–3443.

17. Алексеев, В. Е. Графы и алгоритмы. Структуры данных. Модели вычислений / В. Е. Алексеев. – М. : Интернет-Университет Информационных Технологий ; БИНОМ. Лаборатория знаний, 2012. – 320 с.

18. Diestel, R. Graph Theory / R. Diestel. – New York : Springer, 2000.

19. Дистель, Р. Теория графов : пер. с англ. / Р. Дистель. – Новосибирск : Изд-во Ин-та математики, 2002. – 336 с.

20. von Neumann, J. Theory of Self-Reproducing Automata / J. von Neumann ; ed. and compl. by Arthur W. Burks. – Urbana and London : University of Illinois Press, 1966.

21. фон Нейман, Дж. Теория самовоспроизводящихся автоматов / Дж. фон Нейман ; закон. и отред. А. Бёрксом ; пер. с англ. В. Л. Стефанюка. – Москва : Мир, 1971.

22. Toffoli, T. Cellular Automata Machines: A new environment for modeling / T. Toffoli, N. Margolus. – Cambridge, Massachusetts ; London, England : The MIT Press, 1987.

23. Тоффоли, Т. Машины клеточных автоматов : пер. с англ. / Т. Тоффоли, Н. Марголус. – Москва : Мир, 1991. – 280 с.

24. Li, X. Theory of Practical Cellular Automaton / X. Li, J. Wu, X. Li. – Singapore : Springer Nature, 2018. – 352 p.

25. Essam, J. W. Percolation and Cluster Size / J. W. Essam // Phase Transitions and Critical Phenomena. – New York : Academic Press, 1972. – Vol. 2. – P. 197–270.

26. Weisstein E. W. von Neumann Neighborhood / E. W. Weisstein // MathWorld – A Wolfram Web Resource. URL: https://mathworld.wolfram.com/vonNeumannNeighborhood.html (дата обращения: 14.10.2022).

27. Weisstein E. W. Moore Neighborhood // MathWorld – A Wolfram Web Resource. – URL: https://mathworld.wolfram.com/MooreNeighborhood.html (дата обращения: 14.10.2022).

28. Гордеев, И. И. Программа для нахождения геометрического остова перколяционного кластера методом заливки. Свидетельство о регистрации программы для ЭВМ 2022616981, 18.04.2022. Заявка № 2022616652 от 18.04.2022 / И. И. Гордеев, А. С. Касаткин.

29. Гордеев, И. И. Сравнение алгоритмов Грассбергера и Ахунжанова для нахождения остова перколяционного кластера в задачах перколяции узлов на квадратной решетке / И. И. Гордеев, Н. С. Саенко // Прикаспийский журнал: управление и высокие технологии. – 2022. – № 3 (59). – С. 44–60.

30. G2ii2g. Grassberger-backbone2 // GitHub. – URL: https://github.com/G2ii2g/Grassberger-backbone2 (дата обращения: 14.10.2022).

References

1. Stauffer, D. and Aharony, A. *Introduction to Percolation Theory.* 2nd rev. ed. London, Taylor & Francis, 2003. 180 p.

2. Grimmet, G. *Percolation.* 2nd ed. Berlin ; Heidelberg ; New York ; Barcelona ; Hong Kong ; London ; Milan ; Paris ; Singapore ; Tokyo : Springer, 1999. (Grundlehren der mathematischen Wissenschaften; 321).

3. Tarasevich, Yu. Yu. Percolation: Theory, applications, algorithms : textbook. 2nd ed. Moscow, "LIBROKOM" Booking House, 2012. – 112 p.

4. Sahimi, M. Percolation Phase Transition. *Complex Media and Percolation Theory*. New York, Springer Nature, 2021, pp. 1–9.

5. Roux, S., Hansen, A. A new algorithm to extract the backbone in a random resistor network. *Journal of Physics A: Mathematical and General*, 1987, vol. 20, pp. L1281–L1285.

6. Tarasevich, Yu. Yu., Burmistrov, A. S., Goltseva, V. A., Gordeev, I. I., Serbin, V. I., Sizova, A. A., Vodolazskaya, I. V. and Zholobov, D. A. Identification of current-carrying part of a random resistor network: electrical approaches vs. graph theory algorithms. *Journal of Physics: Conference Series*, 2018, vol. 955, p. 012021. DOI: 10.1088/1742-6596/955/1/012021.

7. Yin, W.-G., Tao, R. Algorithm for finding two-dimensional site percolation backbones. *Physica B: Condensed Matter*, 2000, vol. 279, pp. 84–86.

8. Trobec, R., Stamatovic, B. Analysis and classification of flow-carrying backbones in two-dimensional lattices. *Advances in Engineering Software*, 2017, vol. 103, pp. 38–45.

9. Gordeev, I. I., Ovcharenro, S. S., Sizova, A. A. The determination of the conducting backbone in a two-dimensional lattice by the flooding method. *Caspian Journal: Control and High Technologies*, 2020, no. 1 (49), pp. 94–111. DOI: 10.21672/2074-1707.2020.49.4.094-111.

10. Yin, W.-G., Tao, R. Rapid algorithm for identifying backbones in the two-dimensional percolation model. *International Journal of Modern Physics C*, 2003, vol. 14, pp. 1427–1437. DOI: 10.1142/S0129183103005509.

11. Trobec, R. Index of /~roman/files/tmp/backbone. *Jozef Stefan Institute*. Available at: https://e6.ijs.si/~roman/files/tmp/backbone/ (accessed 09.12.2022).

12. G2ii2g. G2ii2g/backbone-flooding: Program, tests and results for flooding of backbone on square lattice. *GitHub*. Available at: https://github.com/G2ii2g/backbone-flooding (accessed: 09.12.2022).

13. yinweiguo. yinweiguo/FindingBackbone: A program to find the current-carrying backbones in the two-dimensional percolation model. *GitHub*. Available at: https://github.com/yinweiguo/FindingBackbone (accessed 09.12.2022).

14. Tarjan, R. Depth-first search and linear graph algorithms. *SIAM Journal on Computing*, 1972, vol. 1, pp. 146–160.

15. Edwards, R., Goodman, G. J. and Sokal, A. D. Multigrid Method for the Random-Resistor Problem. *Physical Review Letters*, 1988, vol. 61, pp. 1333–1335.

16. Hoshen, J. and Kopelman, R. Percolation and cluster distribution. I. Cluster multiple labeling technique and critical concentration algorithm. *Physical Review B*, 1976, vol. 14, pp. 3438–3443.

17. Alekseev, V. E. *Graphs and algorithms. Data structures. Computation Models.* Moscow, 2012. 320 p.

18. Diestel, R. *Graph Theory*. New York, Springer, 2000.

19. Distel, R. *Graph Theory.* Novosibirsk, 2002. 336 p.

20. von Neumann, J. Burks, W. Arthur (ed. and compl.). *Theory of Self-Reproducing Automata*. Urbana and London, University of Illinois Press, 1966.

21. von Neumann, J. *Theory of self-reproducing automata*. Moscow, Mir Publ., 1971.

22. Toffoli, T., Margolus, N. *Cellular Automata Machines: A new environment for modeling*. Cambridge, Massachusetts ; London, England, The MIT Press, 1987.

23. Toffoli, T., Margolus, N. *Cellular automata machines.* Moscow, 1991. 280 p.

24. Li, X., Wu, J., Li, X. *Theory of Practical Cellular Automaton*. Singapore, Springer Nature, 2018. 352 p.

25. Essam, J. W. Percolation and Cluster Size. *Phase Transitions and Critical Phenomena*, New York, Academic Press, 1972, vol. 2, pp. 197–270.

26. Weisstein, E. W. von Neumann Neighborhood. *MathWorld – A Wolfram Web Resource*. Available at: https://mathworld.wolfram.com/vonNeumannNeighborhood.html (accessed 14.10.2022).

27. Weisstein, E. W. Moore Neighborhood. *MathWorld – A Wolfram Web Resource.* Available at: https://mathworld.wolfram.com/MooreNeighborhood.html (accessed 14.10.2022).

28. Gordeev, I. I., Kasatkin, A. S. *A program for finding the geometric skeleton of a percolation cluster using the filling method.* Certificate of registration of a computer program 2022616981, 04.18.2022. Application no. 2022616652 dated 04.18.2022.

29. Gordeev, I. I., Saenko, N. S. Comparison of Grassberger and Akhunzhanov algorithms for finding the skeleton of a percolation cluster in problems of percolation of nodes on a square lattice. *Caspian Journal: Control and High Technologies*, 2022, no. 3 (59), pp. 44–60.

30. G2ii2g. Grassberger-backbone2. *GitHub*. Available at: https://github.com/G2ii2g/Grassberger-backbone2 (accessed 14.10.2022).

Статья поступила в редакцию 05.10.2023; одобрена после рецензирования 13.10.2023; принята к публикации 17.10.2023.

The article was submitted 05.10.2023; approved after reviewing 13.10.2023; accepted for publication 17.10.2023.

DOI 10.54398/20741707_2023_4_118
УДК 004.056.52:004.732

СИСТЕМА ПРЕДСКАЗАТЕЛЬНОГО МОДЕЛИРОВАНИЯ АКТИВНОСТИ КОНТРАГЕНТОВ НА ОСНОВЕ АНАЛИЗА СТРУКТУРЫ РЕТРОСПЕКТИВНЫХ ДАННЫХ

Кравец Алла Григорьевна, Волгоградский государственный технический университет, 400005, г. Волгоград, пр. им. Ленина, 28.

доктор технических наук, профессор, ORCID: 0000-0003-1675-8652, e-mail: AllaGKravets@yandex.ru

Аль-Мерри Гаис Мохаммед Салех, Волгоградский государственный технический университет, 400005, г. Волгоград, пр. им. Ленина, 28,

преподаватель, ORCID: 0000-0002-9171-6535, e-mail: gaismr2009@mail.ru

В статье описывается разработанная система предсказательного моделирования активности контрагентов на основе анализа структуры ретроспективных данных их предыдущей активности. Для достижения высокой точности предсказаний необходимо собрать достаточно данных об активности контрагентов в прошлом и учесть другие факторы, такие как рыночные условия и изменения законодательства. После этого можно применить различные методы машинного обучения для предсказания будущей активности контрагентов. Результаты данной работы помогут определить, с какими контрагентами следует продолжать сотрудничество, а с какими нет, и принимать соответствующие решения при выборе контрагентов в будущем.

Ключевые слова: система предсказательного моделирования, ретроспективные данные, контрагент, активность, методы машинного обучения

A SYSTEM FOR PREDICTIVE MODELING OF THE ACTIVITY OF COUNTERPARTIES BASED ON THE ANALYSIS OF THE STRUCTURE OF RETROSPECTIVE DATA

Kravets Alla G., Volgograd State Technical University, 28 Lenin Ave., Volgograd, 400005, Russian Federation,

Doct. Sci. (Engineering), Professor, ORCID: 0000-0003-1675-8652, e-mail: AllaGKravets@yandex.ru

Al-Merri G. M. S., Volgograd State Technical University, 28 Lenin Ave., Volgograd, 400005, Russian Federation,

teacher, ORCID: 0000-0002-9171-6535, e-mail: gaismr2009@mail.ru

This article describes a developed system for predictive modeling of the activity of counterparties based on the analysis of the structure of their previous activity retrospective data. To achieve high prediction accuracy, it is necessary to collect sufficient data on the past activity of counterparties and take into account other factors such as market conditions and changes in legislation. After that, various machine learning methods can be applied to predict the future activity of counterparties. The results of this work will help determine which counterparties are trustworthy and which are not, and make appropriate decisions when choosing counterparties in the future.

Keywords: predictive modeling system, retrospective data, counterparty, activity, machine learning methods

ВВЕДЕНИЕ

Контрагенты – это стороны, которые заключают договор или принимают участие в других взаимоотношениях. В бизнесе контрагентами могут быть компании, организации, индивидуальные предприниматели или физические лица. Взаимоотношения могут быть разного характера, например, покупка и продажа товаров или услуг, проведение сделок на финансовых рынках, заключение контрактов на строительство объектов и т. д. Важно иметь информацию о контрагентах, их репутации, финансовом состоянии и других характеристиках для принятия обоснованных решений и минимизации рисков в бизнесе.

Активность контрагента – это совокупность действий и операций, которые проводятся контрагентом, связанных с заключением и исполнением коммерческих сделок.

Предсказательное моделирование – это процесс создания математической модели, которая может использоваться для прогнозирования будущих событий на основе предыдущих данных. Оно используется в различных областях, включая экономику, бизнес и технологии.

Для предсказания активности контрагента могут использоваться различные методы машинного обучения, включая регрессионные модели, нейронные сети, алгоритмы классификации, методы кластеризации и др. [1–3]. Выбор метода зависит от типа данных, которые используются для моделирования, а также от задачи предсказания, которую нужно выполнить. Факторы, влияющие на активность контрагента, могут быть различными и могут включать такие параметры, как финансовые показатели, историю заказов, поведение в социальных сетях и пр. [4–6].

Особенно важно проводить оценку надежности и активности контрагента при заключении государственных контрактов. Однако системы государственных закупок зачастую не предоставляют свои данные даже для исследовательских целей.

Таким образом, целью исследования, представленного в данной статье, является разработка веб-системы предсказательного моделирования активности контрагента (ПМАК) на основе анализа структуры открытых ретроспективных данных (ОРД).

КОНЦЕПЦИЯ СИСТЕМЫ ПРЕДСКАЗАТЕЛЬНОГО МОДЕЛИРОВАНИЯ АКТИВНОСТИ КОНТРАГЕНТА

При разработке системы предсказательного моделирования активности контрагента необходимо учитывать следующие аспекты:

1. Принципы работы рынка. В том числе, какие факторы могут повлиять на контрагентов.
2. Инструменты анализа данных. Оценка рисков контрагента, его финансовой устойчивости и т. д.
3. Алгоритмы машинного обучения. Для разработки системы предсказательного моделирования необходимо использовать алгоритмы машинного обучения, которые будут способны выявлять закономерности и предсказывать активность контрагента.
4. Информационное обеспечение. Разработка системы будет требовать использования специальных информационных технологий для обработки, хранения и анализа больших объемов данных [7].

Таким образом, общая концепция ПМАК может быть представлена в виде следующих компонентов (рис. 1).

1. Серверная подсистема, представляющая собой интеллектуальное ядро. В состав подсистемы входят:

1.1. База ОРД. Обновляемая база с подключением к выбранным источникам ОРД. Реализация базы данных возможна на SQL или NoSQL СУБД. В системе использована СУБД MongoDB, которая обеспечивает гибкость и масштабируемость при хранении документов с различными характеристиками контрагентов, что может быть полезным при работе с большим объемом данных [8].

1.2. Обученная модель с комплексом реализованных методов машинного обучения. Комплекс методов позволяет выбрать лучшую реализацию модели, исходя из структуры и качества ОРД.

2. Веб-интерфейс, предназначенный для задания параметров анализа и визуализации результатов.

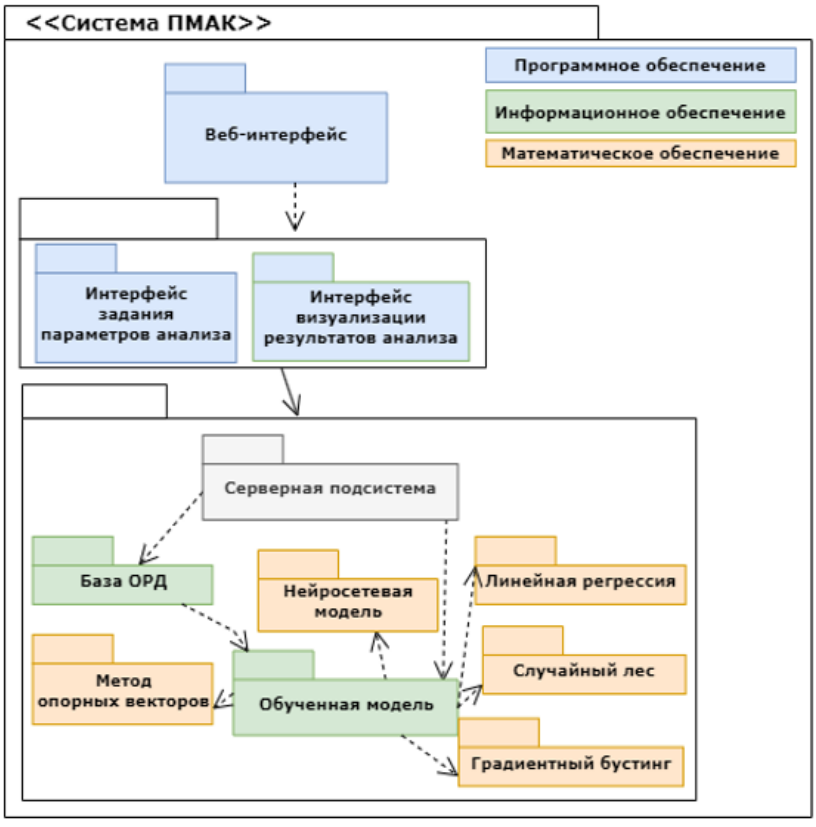


Рисунок 1 – Общая концепция ПМАК

ВЫБОР ДАТАСЕТА И ПРЕДОБРАБОТКА

В ходе исследования были рассмотрены:

1. Государственные датасеты Росреестра[9]:

1.1. Данные о кадастровой стоимости объектов недвижимости в разрезе территориальной принадлежности.

1.2. Данные о ценах регистрируемых сделок (по отчуждению) с объектами недвижимости в разрезе территориальной принадлежности.

1.3. Данные о размере арендной платы за объекты недвижимости в разрезе территориальной принадлежности (в отношении договоров аренды, подлежащих государственной регистрации в соответствии с действующим законодательством) в разрезе территориальной принадлежности.

Данные о контрагентах содержатся только в датасете 1.2, являются узкоспециализированными и не подходят для целей настоящего исследования.

2. Датасет Тинькофф: юридические лица [10]. БД содержит данные о 15000 юридических лицах. Данные обезличены, хорошего качества, содержат описание и «репутацию» юридического лица. Репутационные данные компаний собраны из различных источников: телефонные книжки клиентов (включая негативные и матерные слова в записях), претензии и жалобы, судебные иски, рейтинги в поисковых системах. Датасет похож на синтетический, содержит субъективные оценки и не подходит для целей настоящего исследования.

3. Датасет государственных контрактов [11] проекта OpenRefine. Содержит данные об участии компаний в государственных закупках, а также результаты победы в конкурсных процедурах за 1993–2016 гг. (табл. 1). Данные требуют предобработку, так как содержат пустые строки, данные будущего (например, за 2032 и 2034 гг.). После предобработки датасет можно использовать для обучения моделей.

Таблица 1 – Датасет активности контрагентов в государственных закупках

Характеристика	Параметры
Формат файла	CSV
Количество строк в исходном датасете	6422
Количество строк после очистки	1549
Количество столбцов	9
Описание столбцов	Contract ID – идентификатор контракта Contractor Name – название контрагента Type of Contract – тип контракта Date of Award – дата заключения контракта Start Date – дата начала End Date – дата окончания Total value of Contract – сумма контракта Contract Awarded – признак заключения контракта Unique Project Identifier – идентификатор проекта Agency ID – идентификатор агентства

Также в рамках предобработки произведена замена значения столбца 'Contract Awarded' 'no' на 0, 'yes' на 1. Категориальные признаки кодированы с помощью OneHotEncoder, затем масштабированы с помощью StandardScaler.

АНАЛИЗ ДАННЫХ И ОБУЧЕНИЕ МОДЕЛЕЙ

В ходе анализа статистических данных был использован датасет (табл. 1) для проведения анализа и получения графиков распределения (рис. 2).

В работе использовались пять различных алгоритмов для предсказывания результатов, исходя из входных данных, также было проведено сравнение этих подходов на тестовой выборке. Использовались такие алгоритмы библиотеки Scikit-learn [12], как линейная регрессия (с L1 и L2 регуляризациями), случайный лес, градиентный бустинг, метод опорных векторов, а также нейронные сети. Качество алгоритмов после обучения оценивалось с помощью среднеквадратичного отклонения.

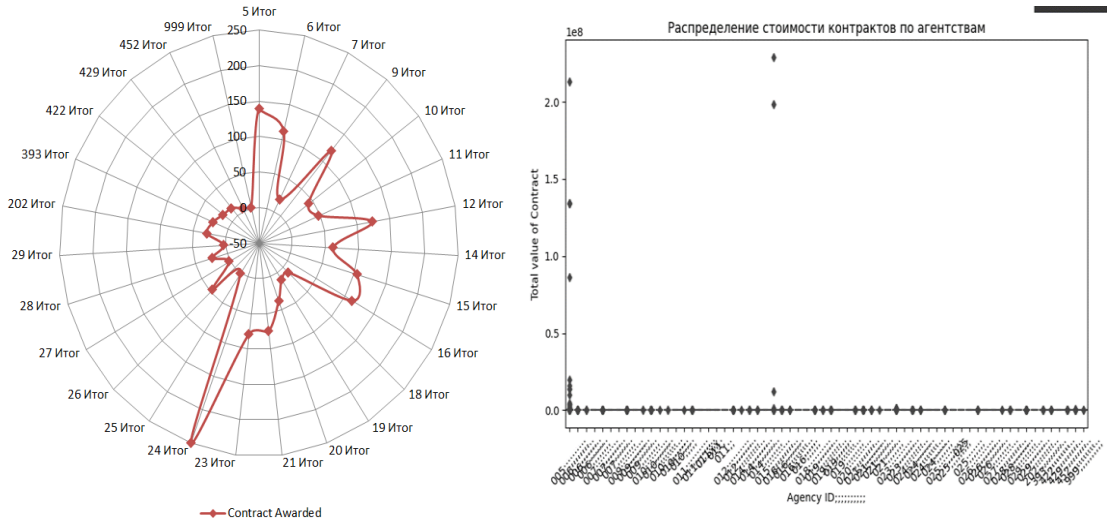


Рисунок 2 – Распределение стоимости контрактов по агентствам

Линейная регрессия

В ходе первого эксперимента обучена модель линейной регрессии [13] без регуляризации и с L1 и L2 регуляризациями со значениями показателя регуляризации α от 10^{-3} до 10^{1}.

Модель линейной регрессии описывается следующей задачей оптимизации:

$$\min(y - w \cdot X - b)^2,$$

где w и b – параметры модели, а (X, y) – данные, оптимизация которых происходит относительно параметров модели. При L1 и L2 регуляризации добавляются следующие члены соответственно:

$$\min[(y - wX - b)^2 + \alpha(|w| + |b|)],$$

$$\min[(y - wX - b)^2 + \alpha(w^2 + b^2)].$$

Модели оказались малочувствительными к значению регуляризации, во всех случаях получали значения коэффициента детерминации на тестовой выборке больше 0,95. Для среднеквадратичного отклонения получили значение ~26, для сравнения дисперсия самих данных составляет ~152. Основная проблема этого подхода при решении данной задачи состоит в том, что модели недообучиваются, то есть значение среднеквадратичного отклонения для обучающей выборки получается больше, чем для тестовой выборки и составляет ~30.

1.1. Случайный лес (СЛ) [14], градиентный бустинг (ГБ) [15], метод опорных векторов (МОВ) [16].

Для нахождения оптимальных параметров данных моделей воспользовались методом кросс-валидации. В процессе кросс-валидации оказалось, что при решении данной задачи эти модели очень чувствительные к разделению данных на обучающую и тестовую выборку, также они очень сильно подвержены к переобучению даже после кросс-валидации. К примеру, для обучающей выборки с помощью градиентного бустинга после кросс-валидации удалось получить отклонение ~2, но для тестовой выборки получено значение ~43. Ситуация с остальными методами аналогичная. Использование данных алгоритмов для решения этой задачи крайне не рекомендуется.

1.2. Нейронная сеть.

Последний используемый метод – это нейронная сеть [17]. Схема нейронной сети, которая показала лучший результат, показана на рисунке 3.

Для обучения нейронной сети определили функцию ошибки как:

$$L(X, y) = E[(y - f_\theta(X))^2],$$

где (X, y) – используемые данные;

$f_\theta(\cdot)$ – нейронная сеть;

θ – параметры нейронной сети.

Обучение нейронной сети происходит с помощью процесса, называемого градиентным спуском, для которого необходимо определить скорость обучения α:

$$\theta_{i+1} = \theta_i - \alpha\nabla_\theta(L(X, y)).$$

Обучение нейронной сети длилось 5000 эпох, в ходе которого скорость обучения с 10^{-3} равномерно уменьшили до 10^{-5}, размер обучающего пакета увеличивали с 10 до 100. В результате получили модель, которая работает намного лучше предыдущих алгоритмов и не подвержена переобучению. Для такой архитектуры нейронной сети получили среднеквадратичное отклонение для тестовой выборки ~14, а для обучающей выборки – ~11.

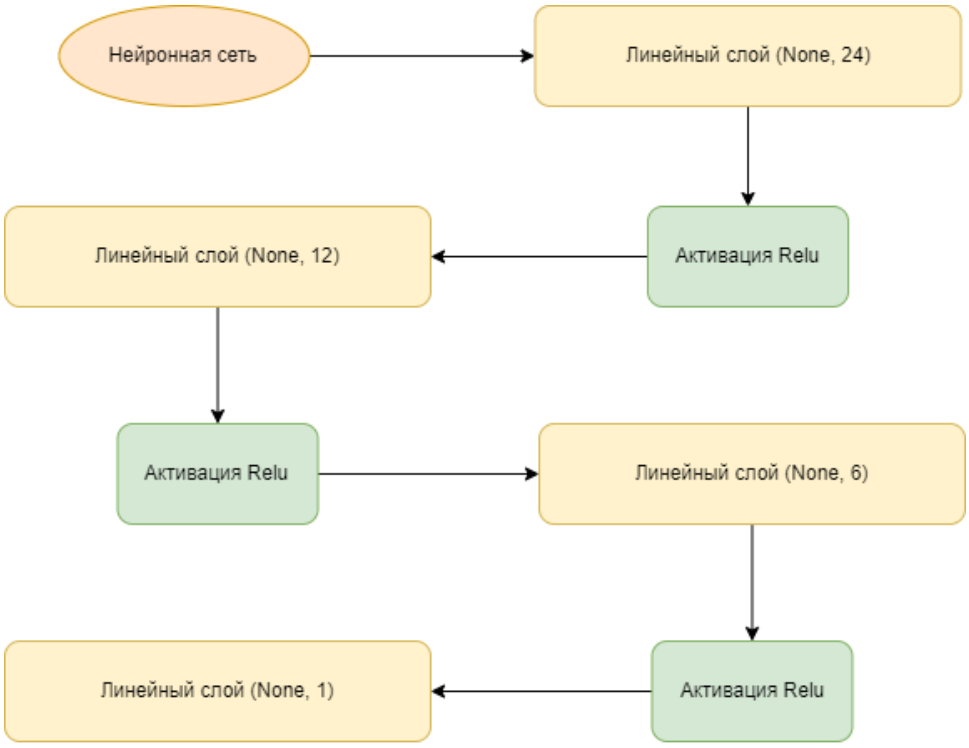


Рисунок 3 – Схема нейронной сети

1.3. Результаты экспериментов.

Результаты экспериментов представлены в таблице 2.

Таблица 2 – Значения среднеквадратичного отклонения для различных моделей

Название модели	Обучающая выборка	Тестовая выборка
Линейная регрессия	31,65	26,33
Линейная регрессия + L1	31,65	26,29
Линейная регрессия + L2	31,65	26,27
СЛ + 500 моделей + 10 глубина	10,30	63,66
СЛ + 500 моделей + 5 глубина	15,29	65,40
СЛ + 100 моделей + 10 глубина	15,46	66,41
СЛ + 100 моделей + 5 глубина	10,06	68,00
ГБ + 100 моделей + 5 глубина	2,18	42,57
ГБ + 100 моделей + 10 глубина	0,71	41,59
МОВ + гауссовое ядро + (C=100) + (γ=1)	149,7	216,8
МОВ + гауссовое ядро + (C=10) + (γ=1)	154,2	217,0
МОВ + гауссовое ядро + (C=100) + (γ=0,1)	149,7	216,8
МОВ + гауссовое ядро + (C=10) + (γ=0,1)	154,2	217,0
Нейронная сеть	10,93	13,78

РАЗРАБОТКА ВЕБ-ИНТЕРФЕЙСА СИСТЕМЫ ПМАК

В разработке веб-интерфейса для системы предсказательного моделирования использовались язык программирования Python и фреймворк Flask. Этот фреймворк обеспечивает возможность создания бекэнда приложения на Python и формирования HTML-страниц для фронтэнда с помощью динамических шаблонов, которые могут быть связаны со статическими CSS и JavaScript. Flask является легковесным и мощным инструментом, который позволяет быстро создавать масштабируемые и производительные веб-приложения, а также предлагает множество библиотек и расширений для улучшения функциональности. Важная роль также отводится каскадным таблицам стилей (CSS), которые позволяют задавать внешний вид HTML-документов и обеспечивать согласованность их внешнего вида [18].

Первая страница представляет из себя стартовый экран (рис. 4), на котором содержится название проекта, а также кнопка перехода к заполнению формы.

После заполнения данных (рис. 5) и отправки запроса в серверную подсистему, они проходят через процесс преобразовательния и передаются модулю, который отвечает за предсказание шансов контракта.

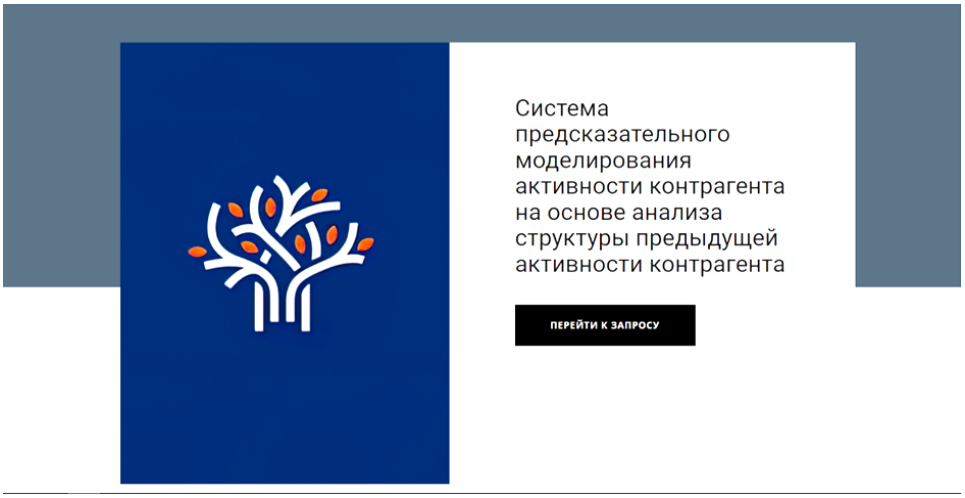


Рисунок 4 – Главный экран

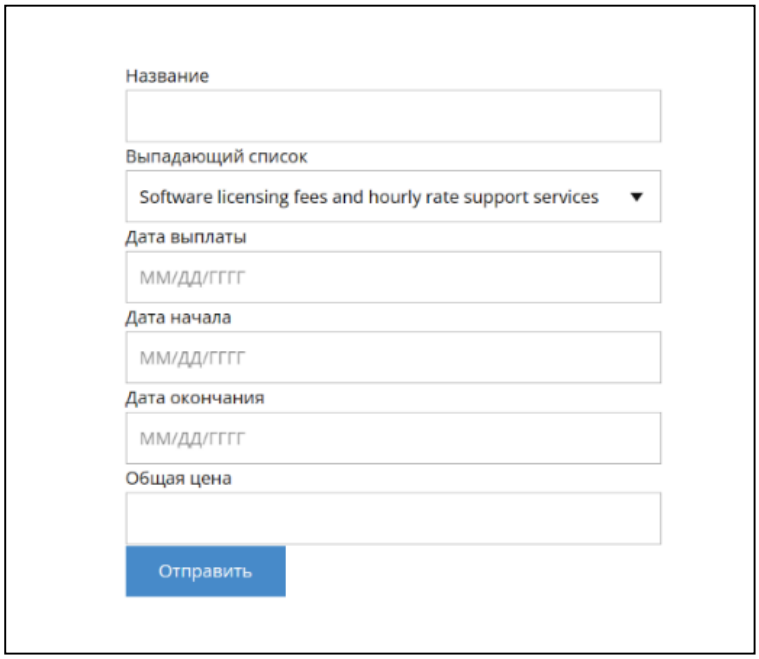


Рисунок 5 – Форма для ввода данных

Модель использует данные из формы для вычисления вероятности успешного заключения котнракта.

ЗАКЛЮЧЕНИЕ

В настоящей работе проведен анализ активности контрагентов, применялись различные методы машинного обучения для предсказания будущей активности контрагентов. В результате был разработан программный продукт, который позволяет вычислить вероятность успешного заключения контракта, он представляет собой инструмент, с помощью которого агентство повышает эффективность своей работы за счет выбора наиболее подходящего варианта для заключения контракта.

Список источников

1. Аль-Мерри, Г. М. С. Управление бизнес-процессами в туристической индустрии на основе методов семантической паутины и онтологического моделирования / Г. М. С. Аль-Мерри, А. Г. Кравец, Н. А. Сальникова // Прикаспийский журнал: управление и высокие технологии. – 2019. – № 1 (45). – С. 73–84. – DOI: 10.21672 /2074-1707.2019.45.1.073-084. – EDN PVBVGD.

2. Zhang, C. et al. A Novel Method for Customer Behavior Prediction Based on Stacked Autoencoder Deep Learning Model / C. Zhang et al. // IEEE Access. – 2018. – Vol. 6. – P. 54735–54745.

3. Al-Merry, G. The development of the searching and analysis methods of the contractors' activity / G. Al-Merry, A. S. Marachovsky // Modern Science and Innovations. – 2017. – № 1 (17). – С. 20–26.

4. Сидоров, А. А. Прогнозирование потребительского спроса на основе анализа и моделирования гидросферных услуг / А. А. Сидоров, А. В. Пантюкова // Гидросфера. – 2015. – № 2.– С. 35–42.

5. Кокуйцев, А. А. Прогнозирование поведения потребителей: методологические подходы и инструменты / А. А. Кокуйцев, О. Ю. Кокуйцева // Менеджмент в России и за рубежом. – 2013. – № 3. – С. 98–107.

6. Голубев, В. А. Предсказание поведения клиента на основе исторических данных / В. А. Голубев, И. Ю. Лапшин // Информатика и ее применения. – 2017. – Т. 11, № 1. – С. 41–47.

7. Kravets, A. G. Tourism cluster enterprises departments resource management based on mobile technologies / A. G. Kravets, A. O. Morozov, K. S. Zadiran, G. Al-Merri, E. A. Trishkina // Communications in Computer and Information Science. – 2019. – Vol. 1083. – P. 218–229.

8. Довбенко, А. В. Хранение данных в NoSQL системах на примере MongoDB / А. В. Довбенко // Вестник науки и образования. – 2015. – № 4 (6). – URL: https://cyberleninka.ru/article/n/hranenie-dannyh-v-nosql-sistemah-na-primere-mongodb (дата обращения: 01.08.2023).

9. Открытая служба. – URL: https://rosreestr.gov.ru/open-service/data-sety-rosreestra/.

10. Национальная олимпиада по анализу данных. Базы данных 2022–2023. – URL: https://dano.hse.ru/data2022 (дата обращения: 01.08.2023).

11. OpenRefine. – URL: https://github.com/OpenRefine/OpenRefine/tree/master (дата обращения: 01.08.2023).

12. Pedregosa, F. Scikit-learn: Machine learning in Python / F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel & E. Duchesnay // Journal of machine learning research. – 2011. – № 12 (Oct). – P. 2825–2830.

13. Осадченко, Р. Анализ данных: поиск золотых крупинок / Р. Осадченко. – URL: https://infostart.ru /public/72928/ (дата обращения: 14.10.2022).

14. Случайный лес. – URL: https://arxiv.org/abs/1407.7502.

15. Градиентный бустинг. – URL: https://arxiv.org/abs/1908.06951.

16. Метод опорных векторов. – URL: https://www.researchgate.net/publication/221621494_Support_Vector_Machines_Theory_and_Applications.

17. Нейронные сети. – URL: http://citebay.com/how-to-cite/pytorch/.

18. Романенко, Е. В. Некоторые вопросы проектирования и реализации распределенной информационной системы "TOUREAST: CRM AI" / Е. В. Романенко, А. Г. Кравец // Прикаспийский журнал: управление и высокие технологии. – 2013. – № 4 (24). – С. 165–176.

References

1. Al-Merri, G. M. S., Kravets, A. G. Salnikova, N. A. Management of business processes in the tourism industry based on the methods of the semantic web and ontological modeling. *Caspian Journal: Control and High Technologies*, 2019, no. 1 (45), pp. 73–84. DOI: 10.21672/2074-1707.2019.45.1.073-084. EDN PVBVGD.

2. Zhang, C. et al. A Novel Method for Customer Behavior Prediction Based on Stacked Autoencoder Deep Learning Model. *IEEE Access*, 2018, vol. 6, pp. 54735–54745.

3. Al-Merry G., Marachovsky A.S. The development of the searching and analysis methods of the contractors' activity. *Modern Science and Innovations*, 2017, no. 1 (17), pp. 20–26.

4. Sidorov, A. A., Pantyukova, A. V. Forecasting consumer demand based on analysis and modeling of hydrosphere services. *Hydrosphere*, 2015, no. 2, pp. 35–42.

5. Kokuytsev, A. A., Kokuytseva, O. Yu. Forecasting consumer behavior: methodological approaches and tools. *Management in Russia and Abroad*, 2013, no. 3, pp. 98–107.

6. Golubev, V. A., Lapshin, I. Yu. Prediction of client behavior based on historical data. *Computer Science and ITS Applications*, 2017, vol. 11, no. 1, pp. 41–47.

7. Kravets, A. G., Morozov, A. O., Zadiran, K. S., Al-Merri, G., Trishkina, E. A. Tourism cluster enterprises departments resource management based on mobile technologies. *Communications in Computer and Information Science*, 2019, vol. 1083, pp. 218–229.

8. Dovbenko, A. V. Data storage in NoSQL systems using the example of MongoDB. *Bulletin of Science and Education*, 2015, no. 4 (6). Available at: https://cyberleninka.ru/article/n/hranenie-dannyh-v-nosql-sistemah-na-primere-mongodb (accessed 08.01.2023).

9. *Open service*. Available at: https://rosreestr.gov.ru/open-service/data-sety-rosreestra/.

10. *National Olympiad in Data Analysis. Databases of 2022–2023*. Available at: https://dano.hse.ru/data2022 (accessed 08.01.2023).

11. *OpenRefine.* Available at: https://github.com/OpenRefine/OpenRefine/tree/master (accessed 01.08.2023)

12. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011, no. 12 (Oct), pp. 2825–2830.

13. Osadchenko, R. *Data analysis: searching for golden grains*. Available at: https://infostart.ru /public/72928/ (accessed 10.14.2022).

14. *Random forest.* Available at: https://arxiv.org/abs/1407.7502.

15. *Gradient boosting.* Available at: https://arxiv.org/abs/1908.06951.

16. *Support vector machine method.* Available at: https://www.researchgate.net/publication /221621494_Support_Vector_Machines_Theory_and_Applications.

17. *Neural networks.* Available at: http://citebay.com/how-to-cite/pytorch/.

18. Romanenko, E. V., Kravets, A. G. Some issues of design and implementation of the distributed information system "TOUREAST: CRM AI". *Caspian Journal: Control and High Technologies*, 2013, no. 4 (24), pp. 165–176.

Статья поступила в редакцию 04.08.2023; одобрена после рецензирования 14.08.2023; принята к публикации 16.08.2023.

The article was submitted 04.08.2023; approved after reviewing 14.08.2023; accepted for publication 16.08.2023.

ПРАВИЛА ДЛЯ АВТОРОВ

1. В журнале публикуются материалы на английском и русском языках по тематике, соответствующей утвержденным для журнала отраслям наук, группам специальностей.

2. В список соавторов работ включаются только те лица, которые внесли творческий вклад в подготовку представленных материалов. Лицам, оказавшим только техническую помощь, можно выразить благодарность в конце статьи. Один человек может быть автором (соавтором) не более чем двух статей в одном номере журнала, причем единственным автором он может быть только в одной статье.

3. Объем публикаций для научных статей должен быть не менее 8 страниц, а количество источников в библиографическом списке (списке литературы) – не менее 10 позиций.

4. Содержание каждой статьи должно включать следующие элементы: УДК; название статьи; сведения об авторах, включая их место работы, должность, адрес электронной почты; аннотацию объемом от 100 до 250 слов, ключевые слова (от 9 до 13); графическую аннотацию, отражающую содержание статьи; название статьи, сведения об авторах, аннотацию и ключевые слова на английском языке (для англоязычных статей – на русском языке); введение – оно должно заканчиваться формулировкой цели работы в явной форме; собственно текст статьи – очень желательна его сегментация на разделы, имеющие содержательные заголовки; выводы или за- ключение (должны соответствовать формулировке цели статьи).

5. Для русскоязычных статей приводится два библиографических списка: на языке оригинала статьи; список с транслитерацией русскоязычных источников на латиницу и (дополнительно) приведением в квадратных скобках переводов названий статей и названий источников на английский язык.

В «русскоязычном» библиографическом списке (списке литературы) порядок следования источников – по алфавиту фамилий авторов (сначала русскоязычные источники, потом иноязычные). На все источники, включенные в библиографический список, должны быть даны ссылки в тексте статьи в квадратных скобках. При необходимости авторы могут указывать номера страниц в источниках, на которые даются ссылки. Приветствуются ссылки на иноязычные источники, а также на материалы, опубликованные ранее в журнале «Прикаспийский журнал: управление и высокие технологии». Однако в последнем случае количество таких ссылок не должно превышать 20 % от общего количества источников, включенных в библиографический список. Для источников, имеющих DOI, целесообразно его указывать. При ссылках на статьи, опубликованные в журнале «Прикаспийский журнал: управление и высокие технологии», целесообразно в конце библиографического описания источника в круглых скобках указывать гиперссылку, указывающую на место размещения статьи на страничке сайта Астраханского государственного университета.

Ссылки в библиографическом списке на материалы, размещенные в интернете, допускаются при соблюдении следующих условий: если у материала, на который дается ссылка, имеется автор и/или название, то они должны быть указаны для этого источника; должен быть приведен полный маршрут доступа к источнику в интернете; должна быть указана дата обращения (доступа) к источнику.

Ограничения по списку литературы: доля самоцитирований для любого из авторов статьи, а также по совокупности всех авторов статьи, не должна превышать 25 %; доля ссылок на статьи с участием одного автора, не являющегося автором (соавтором) статьи, не должна превышать 25 %.

6. Суммарная доля таблиц и иллюстраций в общем объеме представляемой статьи не должна превышать 40 %. Под иллюстрациями понимаются следующие объекты: диаграммы; графики; рисунки; эскизы; фотографии; карты и т.п.

7. Доля оригинального текста в статьях (оцениваемого через систему «Антиплагиат» на сайте www.antiplagiat.ru) должна быть не менее 80 %.

8. Указание на то, что работа финансируется по какому-либо гранту, в рамках Федеральной целевой программы, государственного заказа и пр. дается в виде постраничной сноски после заголовка (названия) работы.

9. В сведения об авторах работ помимо места работы и должности целесообразно включать ORCID автора и гиперссылку на страничку с его личными наукометрическими показателями на сайте www.elibrary.ru. По желанию можно привести также ссылки на странички с наукометрическими показателями на Scopus, в ResearchGate; на личную страничку, размещенную на сайте организации.

10. Основные технические требования к оформлению статей (материалов):

10.1. Текст должен быть расположен по ширине страницы формата А4 с учётом полей (все поля по 2,5 см), набран шрифтом Times New Roman, кегль 12, межстрочный интервал 1,0. В таблицах, подрисуночных надписях допускается уменьшенный шрифт – вплоть до 10 кегля. Альбомная ориентация страниц допускается только в порядке исключения для следующих случаев: широкоформатные таблицы с большим количеством колонок; иллюстрации большого размера, которые не умещаются на странице с книжной ориентацией.

Абзацные отступы одинаковы по всему тексту – 0,75 см. Кавычки («»), скобки ([ ], ( )), маркеры и другие знаки должны быть аналогичными на протяжении всего предоставляемого для публикации материала.

ПРИКАСПИЙСКИЙ ЖУРНАЛ:
управление и высокие технологии

НАУЧНО-ТЕХНИЧЕСКИЙ ЖУРНАЛ

2023
№ 4 (64)

Свидетельство о регистрации средства массовой информации
Федеральной службы по надзору в сфере массовых коммуникаций,
связи и охраны культурного наследия
ПИ № ФС77-31932 от 16 мая 2008 г.

Учредитель
Астраханский государственный университет имени В. Н. Татищева
Российская Федерация, 414056, г. Астрахань, ул. Татищева, 20а

Адрес редакции:
Российская Федерация, 414056, г. Астрахань, ул. Татищева, 20

Адрес издателя:
Российская Федерация, 414056, г. Астрахань, ул. Татищева, 20а

Издание включено в Интернет-каталог
ООО «Агентство «Книга-Сервис» 2022/1

Главный редактор И.М. Ажмухамедов

Редактирование,
компьютерная правка, верстка *Н.Н. Сахно*

Дата выхода в свет 25.12.2023 г.

Цена свободная
Уч.-изд. 11,7. Усл. печ. л. 16,4.
Заказ № 4563. Тираж 500 экз. (первый завод – 22 экз.)

Астраханский государственный университет имени В. Н. Татищева
414056, г. Астрахань, ул. Татищева, 20а
тел. (8512) 24-66-60 (доб. 3; издательско-полиграфический отдел)
E-mail: asupress@yandex.ru
Отпечатано в Астраханской цифровой типографии
414040, г. Астрахань, пл. К. Маркса, 33
тел./факс (8512) 54-00-11, 73-40-40,
E-mail: a-d-t@mail.ru