Учредитель

ФГБОУ ВО «Новосибирский государственный технический университет»

Редакционный совет

Председатель редакционного совета

Вострецов Алексей Геннадьевич, д-р техн. наук, проф., засл. деятель науки РФ, НГТУ, г. Новосибирск

Заместители председателя

Белим Сергей Викторович, д-р физ.-мат. наук, проф., ОмГТУ, г. Омск
Белов Виктор Матвеевич, д-р техн. наук, проф., НГТУ, г. Новосибирск
Котенко Игорь Витальевич, д-р техн. наук, проф., СПИИРАН, г. Санкт-Петербург

Члены редакционного совета

Авдеенко Татьяна Владимировна, д-р техн. наук, проф., НГТУ, г. Новосибирск
Аверченков Владимир Иванович, д-р техн. наук, проф., Брянский ГТУ, г. Брянск
Алгулиев Расим Магомед оглу, д-р техн. наук, проф., академик НАН Республики Азербайджан, ИИТ НАН Республики Азербайджан, г. Баку
Аникин Игорь Вячеславович, д-р техн. наук, доцент, КНИТУ-КАИ, г. Казань
Арутюнян Мариам Евгеньевна, д-р физ.-мат. наук, проф., ИИиАП НАН Республики Армения, г. Ереван
Баранкова Инна Ильинична, д-р техн. наук, доцент, МГТУ им. Г.И. Носова, г. Магнитогорск
Беззатеев Сергей Валентинович, д-р техн. наук, доцент, СПбГУАП, г. Санкт-Петербург
Боранбаев Сейлхан Нарбутинович, д-р техн. наук, проф., Евразийский национальный университет им. Л.Н. Гумилева, г. Нур-Султан, Республика Казахстан
Васильев Владимир Иванович, д-р техн. наук, проф., УГАТУ, г. Уфа
Воевода Александр Александрович, д-р техн. наук, проф., НГТУ, г. Новосибирск
Гатчин Юрий Арменакович, д-р техн. наук, проф., ИТМО, г. Санкт-Петербург
Громов Юрий Юрьевич, д-р техн. наук, проф., Тамбовский ГТУ, г. Тамбов
Иващук Ольга Александровна, д-р техн. наук, проф., НИУ «БелГУ», г. Белгород
Киселёва Тамара Васильевна, д-р техн. наук, проф., СибГИУ, г. Новокузнецк
Кулаков Станислав Матвеевич, д-р техн. наук, проф., СибГИУ, г. Новокузнецк
Кульба Владимир Васильевич, д-р техн. наук, проф., ИПУ РАН, г. Москва
Кытманов Алексей Александрович, д-р физ.-мат. наук, доцент, СФУ, г. Красноярск
Лавлинский Сергей Михайлович, д-р техн. наук, доцент, Институт математики им. С.Л. Соболева СО РАН, г. Новосибирск
Ленский Артем, PhD, ст. науч. сотр., Австралийский национальный университет, г. Канберра
Магазев Алексей Анатольевич, д-р физ.-мат. наук, проф., ОмГТУ, г. Омск
Макарова Елена Анатольевна, д-р техн. наук, проф., УГАТУ, г. Уфа
Мышляев Леонид Павлович, д-р техн. наук, проф., СибГИУ, г. Новокузнецк
Пагано Микеле, д-р, проф., Пизанский университет, г. Пиза, Италия

Пиотровский Дмитрий Леонидович, д-р техн. наук, проф., Средиземноморский Карпасский университет, Турецкая Республика Северного Кипра
Петрунин Юрий Юрьевич, д-р филос. наук, проф., МГУ им. М.В. Ломоносова, г. Москва
Тузиков Александр Васильевич, д-р физ.-мат. наук, проф., чл.-корр. НАН Республики Беларусь, ОИПИ НАН Республики Беларусь, г. Минск
Харин Юрий Семенович, д-р физ.-мат. наук, проф., чл.-корр. НАН Республики Беларусь, БГУ, г. Минск
Ходашинский Илья Александрович, д-р техн. наук, проф., ТУСУР, г. Томск
Шарипов Бахыт Жапарович, д-р пед. наук, проф., Международный университет информационных технологий, г. Алматы, Республика Казахстан
Ячиков Игорь Михайлович, д-р техн. наук, проф., МГТУ им. Г.И. Носова, г. Магнитогорск

Редакция

Главный редактор

Вострецов Алексей Геннадьевич, д-р техн. наук, проф., засл. деятель науки РФ, НГТУ, г. Новосибирск

Заместитель главного редактора

Белов Виктор Матвеевич, д-р техн. наук, проф., НГТУ, г. Новосибирск

Заведующий редакцией

Архипова Анастасия Борисовна, канд. техн. наук, доцент, НГТУ, г. Новосибирск

Журнал зарегистрирован 01.03.2021 Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций. Свидетельство о регистрации ПИ № ФС 77-80320

Адрес издателя и редакции: 630073, г. Новосибирск, пр. К. Маркса, 20.
E-mail: office@publish.nstu.ru и digital-tech-security@mail.ru
Web site: http://publish.nstu.ru и http://journals.nstu.ru/digital-tech-security/

Publisher and editorial office adress: 20 K. Marx Prospekt, Novosibirsk, 630073, Russian Federation

До номера 1 (100) 2021 г. включительно журнал выходил под названием «Сборник научных трудов НГТУ» (ISSN 2307-6879)

16+

© Коллектив авторов, 2022
© Новосибирский государственный технический университет, 2022

БЕЗОПАСНОСТЬ ЦИФРОВЫХ ТЕХНОЛОГИЙ

ISSN 2782-2230

№ 1 (104) **2022**

СОДЕРЖАНИЕ

МЕТОДЫ И СИСТЕМЫ ЗАЩИТЫ ИНФОРМАЦИИ, ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ

Белов В.М., Пивкин Е.Н., Ардаева А.А. Комплексный подход к оцениванию защищенности значимых объектов критической информационной инфраструктуры от несанкционированного доступа 9

Пивкин Е.Н., Ардаева А.А. Анализ методического и технического обеспечения процедур оценивания защищенности значимых объектов критической информационной инфраструктуры от несанкционированного доступа 27

Сидорова Д.Н., Пивкин Е. Н. Алгоритмы и методы кластеризации данных в анализе журналов событий информационной безопасности 41

Хиценко В.Е., Федотов Н.А. Возможности анализа номинативных признаков в задачах информационной безопасности 61

Лемякин В.Е., Архипова А.Б. Исследование применения графов знаний для генерации вопросов аудита информационной безопасности 85

Правила для авторов 99

Выпускающий редактор *И.П. Бровановa*
Корректор *Л.Н. Киншт*
Компьютерная верстка *С.И. Ткачева*

Лицензия № ИД 04303 от 20.03.01. Подписано в печать 25.03.2022. Выход в свет 30.03.2022
Формат 60×84/16. Бумага офсетная. Тираж 300 экз. Уч.-изд. л. 6,04
Печ. л. 6,5. Изд. № 32. Заказ № 124. Цена свободная

Отпечатано в типографии
Новосибирского государственного технического университета
630073, г. Новосибирск, пр. К. Маркса, 20

Editorial board

Novosibirsk State Technical University

Editorial council

Chairman of the editorial council

Vostretsov A.G., Dr. Sc. (Eng.), Novosibirsk State Technical University, Novosibirsk, RF

Deputy chairman

Belim S.V., Dr. Sc. (Phys. & Math.), Omsk State Technical University, Omsk, RF
Belov V.M., Dr. Sc. (Eng.), Novosibirsk State Technical University, Novosibirsk, RF
Kotenko I.V., Dr. Sc. (Eng.), St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, RF

The members of the editorial council

Avdeenko T.V., Dr. Sc. (Eng.), Novosibirsk State Technical University, Novosibirsk, RF
Averchenkov V.I., Dr. Sc. (Eng.), Bryansk State Technical University, Bryansk, RF
Alguliyev R.M.o., Dr. Sc. (Eng.), Azerbaijan National Academy of Sciences, Institute of Information Technology, Baku, AZE
Anikin I.V., Dr. Sc. (Eng.), Kazan National Research Technical University named after A.N. Tupolev – KAI, Kazan, RF
Haroutunian M.E., Dr. Sc. (Phys. & Math.), Institute for Informatics and Automation Problems of NAS RA, Yerevan, ARM
Barankova I.I., Dr. Sc. (Eng.), Nosov Magnitogorsk State Technical University, Magnitogorsk, RF
Bezzateev S.V., Dr. Sc. (Eng.), Saint Petersburg State University of Aerospace Instrumentation, St. Petersburg, RF
Boranbaev S.N., Dr. Sc. (Eng.), L.N. Gumilyov Eurasian National University, Nur-Sultan, KZ
Vasil'ev V.I., Dr. Sc. (Eng.), Ufa State Aviation Technical University, UFA, RF
Voevoda A.A., Dr. Sc. (Eng.), Novosibirsk State Technical University, Novosibirsk, RF
Gatchin Yu.A., Dr. Sc. (Eng.), National Research University of Information Technologies, Mechanics and Optics, St. Petersburg, RF
Gromov Yu.Yu., Dr. Sc. (Eng.), Tambov State Technical University, Tambov, RF
Ivashhuk O.A., Dr. Sc. (Eng.), Belgorod State National Research University, Belgorod, RF
Kiseljova T.V., Dr. Sc. (Eng.), Siberian State Industrial University, Novokuznetsk, RF
Kulakov S.M., Dr. Sc. (Eng.), Siberian State Industrial University, Novokuznetsk, RF
Kul'ba V.V., Dr. Sc. (Eng.), V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Moscow, RF
Kytmanov A.A., Dr. Sc. (Phys. & Math.), Siberian Federal University, Krasnoyarsk, RF
Lavlinskij S.M., Dr. Sc. (Eng.), Sobolev Institute of Mathematics of Russian Academy of Sciences, Novosibirsk, RF
Lenskij A., PhD, Australian National University, Canberra, AUS
Magazev A.A., Dr. Sc. (Phys. & Math.), Omsk State Technical University, Omsk, RF
Makarova E.A., Dr. Sc. (Eng.), Ufa State Aviation Technical University, Ufa, RF

Myshljaev L.P., Dr. Sc. (Eng.), Siberian State Industrial University, Novokuznetsk, RF
Pagano M., Dr. Sc., University of Pisa, Pisa, IT
Piotrovskij D.L., Dr. Sc. (Eng.), University of Mediterranean Karpasia, Turkish Republic of Northern Cyprus, CYP
Petrunin Yu.Yu., Dr. Sc. (Philos.), Lomonosov Moscow State University, Moscow, RF
Tuzikov A.V., Corresponding Member, National Academy of Sciences of Republic Belarus, Dr. Sc. (Phys. & Math.), United Institute of Informatics Problems, Minsk, BLR
Harin Yu.S., Corresponding Member, National Academy of Sciences of Republic Belarus, Dr. Sc. (Phys. & Math.), Belarusian State University, Minsk, BLR
Hodashinskij I.A., Dr. Sc. (Eng.), Tomsk State University of Control Systems and Radioelectronics, Tomsk, RF
Sharipov B.Zh., Dr. Sc. (Ped.), International University of Information Technology, Almaty, KZ
Jachikov I.M., Dr. Sc. (Eng.), Nosov Magnitogorsk State Technical University, Magnitogorsk, RF

Editorial office

Chief editor

Vostretsov A.G., Dr. Sc. (Eng.), Novosibirsk State Technical University, Novosibirsk, RF

Deputy chief editor

Belov V.M., Dr. Sc. (Eng.), Novosibirsk State Technical University, Novosibirsk, RF

Head of the editorial office

Arhipova A.B., Candidate of Science (Eng.), Novosibirsk State Technical University, Novosibirsk, RF

Publisher and editorial adress: 20 K. Marx Prospekt, Novosibirsk, 630073, Russian Federation
E-mail: office@publish.nstu.ru, digital-tech-security@mail.ru
Web site: http://publish.nstu.ru, http://journals.nstu.ru/digital-tech-security/

© Authors, 2022
© Novosibirsk State Technical University, 2022

DIGITAL TECHNOLOGY SECURITY

ISSN 2782-2230

№ 1 (104) **2022**

CONTENTS

METHODS AND SYSTEMS OF INFORMATION PROTECTION, INFORMATION SECURITY

Belov V.M., Pivkin E.N., Ardaeva A.A. An integrated approach to assessing the security of significant objects of critical information infrastructure from unauthorized access 9

Pivkin E.N., Ardaeva A.A. Analysis of methodological and technical support of procedures for assessing the security of significant objects of critical information infrastructure from unauthorized access 27

Sidorova D.N., Pivkin E.N. Algorithms and methods of data clustering in the analysis of information security event logs 41

Khitsenko V.E., Fedotov N.A. Possibilities of analysis of nominative signs in tasks of information security 61

Lemyakin A.V., Arkhipova A.B. Research of application of knowledge graphs to generate information security audit questions 85

Rules for authors 99

Editor *L.N. Kinsht*
Publishing Editor *I.P. Brovanova*
Computer imposition *S.I. Tkacheva*

License № ID 04303 from 20.03.01. Signed in print March 25, 2022
Date of publication March 30, 2022. Format 60 × 84 1/16
Offset Paper. Circulation is 300 copies. Educational-ed. liter. 6,04. printed pages 6,5
Publishing number 32. Order number 124

It is printed in printing house of Novosibirsk State Technical University
630073, Novosibirsk, 20 K. Marx Prospekt

БЕЗОПАСНОСТЬ ЦИФРОВЫХ ТЕХНОЛОГИЙ. – 2022. – № 1 (104). – 9–26

МЕТОДЫ И СИСТЕМЫ ЗАЩИТЫ ИНФОРМАЦИИ, ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ

УДК 004.9 DOI: 10.17212/2782-2230-2022-1-9-26

КОМПЛЕКСНЫЙ ПОДХОД К ОЦЕНИВАНИЮ ЗАЩИЩЕННОСТИ ЗНАЧИМЫХ ОБЪЕКТОВ КРИТИЧЕСКОЙ ИНФОРМАЦИОННОЙ ИНФРАСТРУКТУРЫ ОТ НЕСАНКЦИОНИРОВАННОГО ДОСТУПА*

В.М. БЕЛОВ[1], Е.Н. ПИВКИН[2], А.А. АРДАЕВА[3]

[1] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, доктор технических наук, профессор кафедры защиты информации. E-mail: vmbelov@mail.ru*
[2] *105066, РФ, г. Москва, ул. Новорязанская, 31/7, к. 2, ПАО АКБ «Связь-Банк», кандидат технических наук. E-mail: evpiv@yandex.ru*
[3] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, магистрант кафедры вычислительной техники. E-mail: nastya.ardaeva@yandex.ru*

Количество событий кибербезопасности в современном мире значительно возросло, из них значительное число приходится на объекты критической информационной инфраструктуры. В настоящей работе рассматриваются основные требования, предъявляемые к моделям оценивания защищенности значимых объектов критической информационной инфраструктуры от несанкционированного доступа, классификация неоднозначной исходной информации, обобщенный алгоритм оценивания уровня защищенности значимых объектов критической информационной инфраструктуры от несанкционированного доступа, нечеткая модель оценивания уровня защищенности с использованием балльной и лингвистической шкал. Определяется порядок проведения оценивания защищенности значимых объектов критической информационной инфраструктуры от несанкционированного доступа. Среди основных требований к моделям оценивания защищенности значимых объектов критической информационной инфраструктуры от несанкционированного доступа выделяют следующие: универсальность, расширяемость, формализуемость, простота, многофакторность. Обобщенный алгоритм оценивания уровня защищенности значимых объектов критической информационной инфраструктуры от несанкционированного доступа включает четыре вычислительных блока: в блоке 1 проводится сбор и первичная обработка информации; в блоке 2 используются вычисления по алгоритмам нечеткого оценивания с лингвистической и балльной шкалами; в блоке 3 оценивается эффективность работы сотрудников по информационной

* Статья получена 02 февраля 2022 г.

безопасности; в блоке 4 осуществляется прогнозирование уровней защищенности значимых объектов критической информационной инфраструктуры от несанкционированного доступа с помощью нечетких временных рядов; в блоке 5 полученные на предыдущих этапах вычислений оценки исследуются с использованием методик обработки результатов данных и делаются соответствующие выводы для принятия решений.

Ключевые слова: модели оценивания защищенности, классификация исходной информации, алгоритм оценивания уровня защищенности, нечеткая модель оценивания уровня защищенности, балльная шкала, лингвистическая шкала, порядок оценивания защищенности, несанкционированный доступ

1. ТРЕБОВАНИЯ, ПРЕДЪЯВЛЯЕМЫЕ К МОДЕЛЯМ ОЦЕНИВАНИЯ ЗАЩИЩЕННОСТИ ЗО КИИ

Выделяют следующие основные требования к оцениванию уровней защищенности значимых объектов критической информационной инфраструктуры (ЗО КИИ) от несанкционированного доступа [1–14]:

- универсальность – позволяет использовать модель для представления различных типов ЗО КИИ независимо от источника угроз, объекта и средств защиты информации (СЗИ);
- расширяемость – обеспечивает возможность добавления новых характеристик защищенности к модели;
- формализуемость – признак, указывающий на возможность применения математических инструментов для описания параметров модели;
- простота – параметр, который позволяет пользователю легко воспринимать структуру и способы реализации моделируемой защищенности ЗО КИИ;
- многофакторность – позволяет учитывать основные параметры моделируемой защищенности ЗО КИИ.

Исходя из перечисленных требований к оцениванию уровней защищенности ЗО КИИ, были выделены следующие критерии, по которым можно определить эффективность модели оценивания уровней защищенности ЗО КИИ:

- возможность проведения оценивания уровней защищенности ЗО КИИ;
- возможность проведения оценивания уровней защищенности новых ЗО КИИ, описание которых неизвестно;
- возможность идентифицировать процесс вывода результатов применения модели, т. е. возможность более точно определить причины, по которым было принято решение об оценке уровня защищенности ЗО КИИ;
- расширяемость – свойство, позволяющее добавлять в модель дополнительные параметры, которые способны определить различные уровни защищенности ЗО КИИ.

2. КЛАССИФИКАЦИЯ НЕОДНОЗНАЧНОЙ ИСХОДНОЙ ИНФОРМАЦИИ

Проблема выбора альтернатив (принятия решений) – задача, в которой решения принимают в условиях, когда поставленные цели, имеющие ограничения и следствия, порождаются возможными, точно неизвестными действиями. Рассмотрение разнообразных неясных, неопределенных и неточных явлений, событий или фактов, а также связей между объектами и операциями показывает существование различных классов неясности или неопределенности, которые не всегда связаны со случайностью или нечеткостью [1–14].

Имеются признаки необходимости применения нечеткой модели (НМ):

– информация о системе имеет различное качество, а значения параметров оцениваются по разным шкалам;

– нечеткость и неясность выбора или описания границы системы или ее отдельных состояний, а также входных и выходных воздействий;

– для описания функционирования систем в виде эвристических предпочтений используется конструкция естественного языка в виде правил «если – то».

Нечеткая модель при наличии обучающей выборки позволяет аппроксимировать функции или измеряемые данные с любой желаемой точностью [1–14].

Наиболее успешно математическая теория нечеткости представлена нечеткой логикой. Альтернативная теория множеств (АТМ) является уже другой глубокой математической теорией, которая адресована, помимо прочего, также и нечеткости [15–17].

Сложность выбора показателей, позволяющих дать адекватную оценку уровню защищенности ЗО КИИ, определяют:

– необходимостью контроля большого количества средств и объектов защиты, а также мероприятий, направленных на сохранение основных свойств информации (конфиденциальность, целостность, доступность);

– нечеткостью проявления внешних воздействий и внутренних изменений (угрозы, КУИ, нарушители) на ЗО КИИ и системах ее защиты;

– отсутствием аналогов показателей, учитывающих специфику ЗО КИИ и особенности ее функционирования;

– необходимостью получения не только качественной, но и количественной оценки уровня защищенности ЗО КИИ.

Особый интерес представляют такие модели и системы, которые дают возможность эффективно проводить экспертизу (с учетом качественной оценки) и осуществлять выбор наилучшего варианта компьютерной системы, используя интегрированные значения уровней защищенности ЗО КИИ. Такие модели позволяют оценивать уровни защищенности ЗО КИИ и выбирать аль-

тернативы на основе экспертных оценок, эталонных параметров, нечетко определенных исходных данных разработчиков и экспертов, которые представлены как в количественной, так и в лингвистической форме.

3. МЕТОДЫ ОЦЕНИВАНИЯ СОСТОЯНИЯ ЗАЩИЩЕННОСТИ ЗО КИИ

Методы оценивания состояния защищенности подразделяют на количественные и качественные (рис. 1) [1–14]. В количественном подходе анализируют степень риска, так как эффективность СЗИ относительно потерь нельзя оценить, если нет количественных показателей степени защищенности.

Качественный подход эффективно применяют в случае, когда потенциальная потеря незаметна, а риск нельзя выразить в денежном эквиваленте. При таком подходе результаты защищенности выражают в лингвистической форме [1–14].

Рис. 1. Методы анализа защищенности ЗО КИИ

Fig. 1. Methods for analyzing the security of the CII

Решая задачи в области ИБ, исследуют корректность работы таких систем в наиболее интересных для специалистов условиях функционирования с заведомо прогнозируемыми значениями параметров, которые адекватно отображают свойства системы.

4. ОБОБЩЕННЫЙ АЛГОРИТМ ОЦЕНИВАНИЯ УРОВНЯ ЗАЩИЩЕННОСТИ

Средства анализа защищенности ЗО КИИ, основанные на нечеткой модели, определяют текущий уровень защищенности ЗО КИИ и позволяют осуществлять прогнозирование динамики его дальнейшего изменения [1–14].

Все перечисленные операции выполняют люди, обладающие совершенными знаниями в области информационной безопасности (ИБ). Но бывают и случаи неполноты знаний, субъективизма предпочтений, неопределенности целей и критериев и т. д. Это особенно заметно при проведении внутреннего аудита (самооценки).

Обобщенный алгоритм оценивания уровня защищенности представляет собой последовательность блоков реализации (рис. 2).

Блок 1 является одним из самых сложных, поскольку требуется произвести сбор, обработку информации, учитывая следующие ограничения:

– нечеткие неполные данные: преобладание качественных данных при небольшом числе количественных, отсутствие предыстории, невоспроизводимость условий угроз, КУИ и атак, ограничения по достижимости точности результата оценки, невозможность сбора исчерпывающих данных;

– нечеткие неполные знания: неадекватность оценки ситуации различными сотрудниками (особенно руководства, отделов информационных технологий, информационной безопасности и других отделов).

Оценивание уровней защищенности ЗО КИИ (блок 2) предполагает использование НМ с лингвистической (НМЛШ) или балльной (НМБШ) шкалами [1–14].

В зависимости от способа представления шкалы, формата шкалы и необходимого быстродействия выбирают одну из НМ.

Различие между полученными оценкой и самооценкой можно рассматривать как элемент для определения эффективности работы сотрудников отдела ИБ (ОИБ). Проведенное обследование является исходной информацией для определения эффективности работы сотрудников по ИБ (блок 3). Причем здесь используют как количественные показатели, так и качественные. В нашей работе блок 3 не анализируется, так как предполагает изучение персональных данных сотрудников ЗО КИИ, на которое требуется специальное разрешение организации.

На основе определенных оценок текущего уровня защищенности и определении эффективности работы сотрудников по ИБ осуществляют прогнозные оценки (блок 4). Прогнозирование осуществляют с использованием модели нечетких временных рядов. В нашей работе блок 4 не используется, так как предполагает наличие персональных данных сотрудников ЗО КИИ из блока 3.

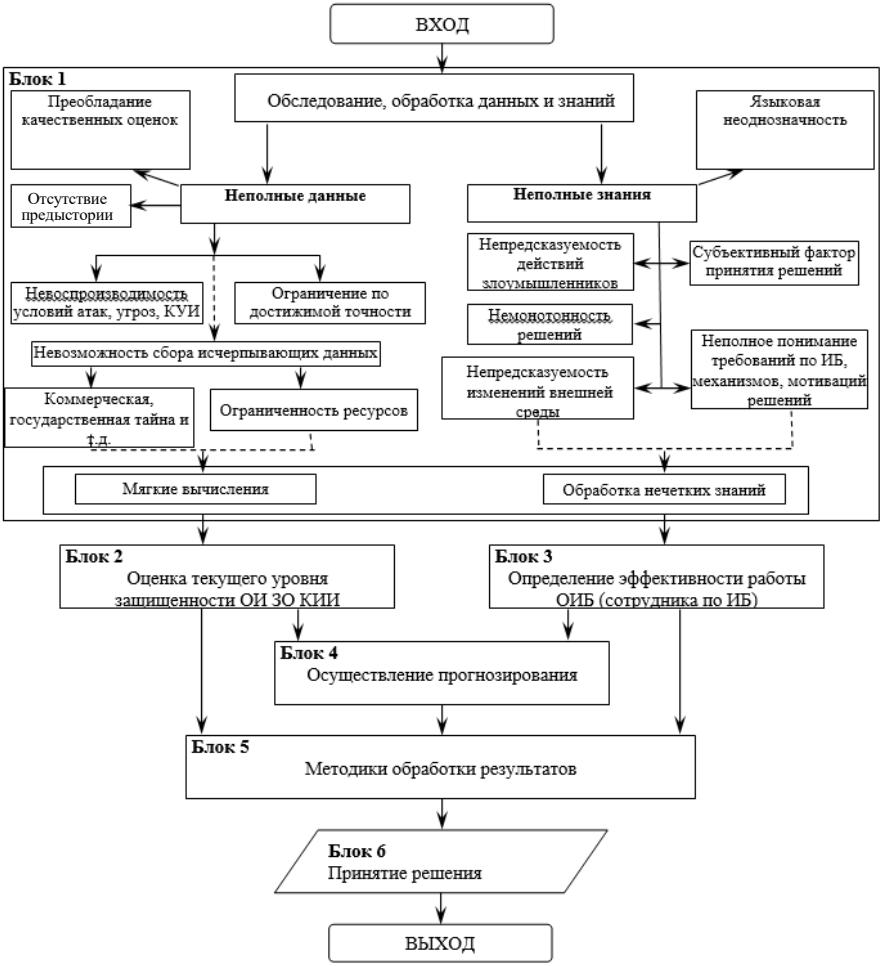


Рис. 2. Обобщенный алгоритм нечеткой модели (комплексный подход)

Fig. 2. Generalized fuzzy model algorithm (complex approach)

Полученные результаты оценок текущего уровня защищенности ЗО КИИ (на основе оценки и самооценки), эффективности работы сотрудников по ИБ, а также прогнозные оценки исследуют с использованием методик обработки

результатов (блок 5) и на их основании делают соответствующие выводы и принимают решения.

5. НЕЧЕТКАЯ МОДЕЛЬ ОЦЕНИВАНИЯ УРОВНЯ ЗАЩИЩЕННОСТИ С ИСПОЛЬЗОВАНИЕМ БАЛЛЬНОЙ ШКАЛЫ

В НМБШ (рис. 3) [1–14] определяет уровень защищенности ЗО КИИ нечеткими термами T_i с использованием лингвистических переменных (ЛП).

Сотрудник по ИБ (ОИБ) отвечает на предварительно ранжированные вопросы. Эти вопросы захватывают всю программу внутреннего аудита ЗО КИИ и ранжируются в соответствии с важностью по *N*-балльной шкале. Через определенный промежуток времени сотрудники просматривают вопросы и варьируют диапазон шкалы $[X_j, \overline{X_j}]$ в зависимости от актуальных угроз модели угроз и нарушителя безопасности информации и каналов утечки информации (КУИ) ЗО КИИ, нормативных документов по ИБ.

Рис. 3. Схема определения уровня защищенности ЗО КИИ, основанная на НМБШ

Fig. 3. Scheme for determining the level of security of the CA of the CII, based on the NMBSH

Диапазон $[X_j, \overline{X_j}]$ $(X_j = 0,\ \overline{X} = N_j)$ изменений параметра X_j^* (где j = [1, n], N_j – максимально возможное количество баллов для каждого вопроса, n – число вопросов) отображается на универсальное множество эталонных нечетких чисел (НЧ) $X_{\text{уз}} = [0, L-1]$ (L – количество эталонов), для которого фиксированное значение $X_j^* \in [\underline{X}_j, \overline{X}_j]$ переводится в соответствующий элемент $X_{\text{уз}\,j}^* \in [0, L-1]$ по формуле

$$X_{\text{уз}\,j}^* = (L-1)\frac{X_j^* - \underline{X}_j}{\overline{X}_j - \underline{X}_j}. \tag{1}$$

Далее задается функция принадлежности (ФП) $\mu_i^j\left(X_{\text{уз}\,j}^*\right)$ (где $i = [1, L]$) нечеткого терма с i-м номером в треугольном виде [17–30] с учетом коэффициентов важности (КВ) PN_j, $j = [1, n]$, вычисленных по оценкам проверяющих ЗО КИИ для каждого компонента вышеприведенной программы аудиторской проверки внутреннего аудита ЗО КИИ.

На завершающей стадии определяют показатель уровня защищенности ЗО КИИ по следующему логическому выражению:

$$\mu_S\left(X_j^*\right) = \bigvee_{i=1}^{L} \bigwedge_{j=1}^{n} \mu_i^j, \tag{2}$$

где $i = [1, L]$ – номер терма из базового терм-множества T; $j = [1, n]$ – номер компонента экспертного запроса.

6. НЕЧЕТКАЯ МОДЕЛЬ ОЦЕНИВАНИЯ УРОВНЯ ЗАЩИЩЕННОСТИ С ИСПОЛЬЗОВАНИЕМ ЛИНГВИСТИЧЕСКОЙ ШКАЛЫ

Определение состояния защищенности ЗО КИИ в соответствии с НМЛШ (рис. 4) [1–14] реализуют по результатам опроса сотрудников ОИБ (сотрудника по ИБ), разделы (вопросы) предварительно ранжируют через определение КВ: P_i ($i = [1, n]$, n – количество разделов).

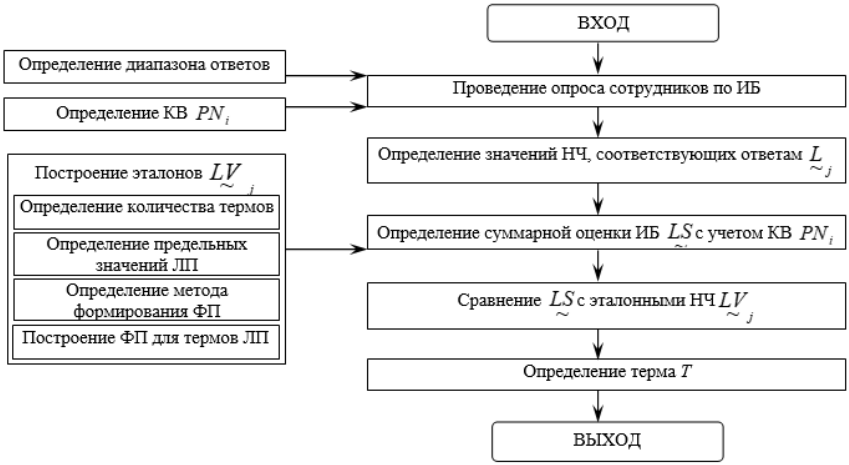


Рис. 4. Схема определения уровня защищенности ЗО КИИ, основанная на НМЛШ

Fig. 4. Scheme for determining the level of security of the CII, based on the NMLSH

Для достижения этой цели используется метод оценивания, основанный на преобразованной матрице $A' = \left(a'_{vw}\right)$, которая получается из матрицы парных сравнений $A = (a_{ij})$ (см. таблицу). Элементы преобразованной матрицы определяются так:

$$\begin{cases} 100/(a_{ij}+1)a_{ij}, & i<j:\ v = i,\ w = j, \\ a'_{vw} = 1, & i<j:\ v = w = i = j, \\ 100/(a_{ij}+1), & i<j:\ v = j,\ w = i, \end{cases} \tag{3}$$

где $i = j = [1, n]$, n – количество разделов (вопросов) программы проверки.

Значения КВ $(P_i, i = [1, n])$ для каждого из вопросов и разделов программы рассчитываются по формуле

$$P_i = \sum_{i=1}^{n} a_{ij} \quad (i \neq j). \tag{4}$$

Шкала для построения матрицы суждений

Scale for constructing a matrix of judgments

Оценка значимости	Качественная оценка	Примечание
1	Одинаковая значимость	Альтернативы имеют одинаковый ранг
3	Слабое преимущество	Преимущество одной альтернативы перед другой неубедительное
5	Сильное преимущество	Имеются достоверные доказательства значительного преимущества одной альтернативы
7	Очевидное преимущество	Имеются веские доказательства в пользу одной альтернативы
9	Абсолютное преимущество	Имеются веские доказательства в пользу преимущества одной альтернативы перед другой с наибольшей степенью убедительности
2, 4, 6, 8	Промежуточные значения	Используются, если требуется компромисс

После определения КВ осуществляется нормализация по выражению

$$PN_i = P_i \Big/ \left(\sum_{i=1}^{n} P_i \right), \tag{5}$$

для того чтобы выполнялось условие

$$\sum_{i=1}^{n} PN_i = 1 . \tag{6}$$

Помимо ранжирования вопросов и разделов по степени важности сотрудники, осуществляющие проверку, строят нечеткие эталоны, отображающие лингвистическую переменную «Уровень защищенности», которая является образцом для сравнения НЧ.

Модель НМЛШ (рис. 4) предполагает, что группа из N определенного количества сотрудников ОИБ отвечает на n вопросов (n-компонентный

экспертный запрос). По ответам сотрудников ОИБ формируется НЧ $Z_t(t=[1,N])$, которому ставят в соответствие одно из эталонных. Значения НЧ, соответствующие оценке ответов всей группы сотрудников ОИБ на j-й раздел (вопрос) $(j=[1,n])$, определяется по формуле

$$\underset{\sim}{L}_j = \left(\sum_{t=1}^{N} \underset{\sim}{Z}_t \right) \Big/ N, \tag{7}$$

где $\sum\limits_{\sim}$ – нечеткое сложение, выполненное по одному из методов реализации операций нечеткой арифметики [1–14].

Суммарную оценку защищенности определяют с учетом ранее вычисленных КВ:

$$\underset{\sim}{LS} = \left(\sum_{j=1}^{n} PN_i \underset{\sim}{L}_j \right). \tag{8}$$

Образованное $\underset{\sim}{LS}$ сравнивают с эталонными НЧ, для чего используют α-уровневое расстояние (АУР) [16]:

$$d(\underset{\sim}{LS}, \underset{\sim}{LV}_j) = \left(\sum_{j=1}^{k} \sum_{i=1}^{m} |x_i - y_j| \right) \Big/ k, \tag{9}$$
$$(\forall X_{\text{уз}y} \geq \alpha)$$

где α – заданное значение α-уровня $(0 \leq \alpha \leq 1)$; x_i и y_i – носители полученного и эталонного НЧ $\underset{\sim}{LS}$ и $\underset{\sim}{LV}_j$; m – количество компонентов НЧ $\underset{\sim}{LS}$; k – количество компонентов НЧ $\underset{\sim}{LV}_j$ с ФП $\mu_y \geq \alpha$.

Критерием соответствия $\underset{\sim}{LS}$ одному из эталонных НЧ считают минимальное АУР, которое и определяет уровень защищенности ЗО КИИ:

$$d\min_i = \bigwedge_{j=1}^{k} d\left(\underset{\sim}{LS}, \underset{\sim}{LV}_j\right), \tag{10}$$

где $\underset{\sim}{LV}_j$ – эталонные НЧ.

7. ПОРЯДОК ПРОВЕДЕНИЯ ОЦЕНИВАНИЯ ЗАЩИЩЕННОСТИ ЗО КИИ

Для проведения оценивания защищенности [1–14] ЗО КИИ предполагается соблюдение следующих этапов.

1. Необходимо провести осмотр ЗО КИИ, анализ документов по ИБ, их актуальность, достоверность, соответствие положений дел в требованиях документов по ИБ России, опрос сотрудников по ИБ и аудит ИБ с целью своевременного выявления, оценивания и прогнозирования источников угроз ИБ, причин и условий, способствующих нанесению ущерба интересам России, нарушению нормального функционирования и развития ИТКС России.
2. Составить вопросы по различным направлениям ИБ. Определить коэффициенты важности, используя метод относительного ранжирования, или руководствоваться требованиями на законодательном уровне.
3. Сформировать эталонные значения уровней защищенности ЗО КИИ методом лингвистических термов.
4. При проведении аудиторской проверки внутреннего аудита организации провести опрос сотрудников по ИБ, в случае необходимости – опрос специалистов других подразделений.
5. Определить суммарную оценку уровня защищенности и провести сравнение со всеми эталонными значениями ЗО КИИ.

ЗАКЛЮЧЕНИЕ

Определение оценки уровня защищенности ЗО КИИ играет важную роль в формировании комплекса практических мер по ее реализации. Но, к сожалению, принятие решения с помощью классических методов и моделей определенно не предполагает желаемого эффекта.

В настоящей работе были рассмотрены основные требования, предъявляемые к моделям оценивания защищенности ЗО КИИ, классификация неоднозначной исходной информации, обобщенный алгоритм оценивания уровня защищенности ЗО КИИ, нечеткая модель оценивания уровня защищенности с использованием балльной и лингвистической шкал, а также определен порядок проведения оценивания защищенности ЗО КИИ.

СПИСОК ЛИТЕРАТУРЫ

1. *Белов В.М., Пивкин Е.Н.* Анализ проблем создания модели комплексной системы защиты информации в региональных налоговых органах // Актуальные проблемы безопасности информационных технологий: материалы I Международной заочной научно-практической конференции / Сибирский государственный аэрокосмический университет. – Красноярск, 2007. – С. 19–21.

2. *Белов В.М., Пивкин Е.Н.* Информационные потоки и способы их формализации с использованием аппарата нечетких множеств в региональных налоговых органах // Инфокоммуникационные системы и технологии: проблемы и перспективы / под ред. А.В. Бабкина. – СПб., 2007. – С. 336–339.

3. *Белов В.М., Пивкин Е.Н., Прокопец В.Д.* Основные математические модели и методы при построении систем защиты информации // Инфокоммуникационные системы и технологии: проблемы и перспективы / под ред. А.В. Бабкина. – СПб., 2007. – С. 300–331.

4. *Белов В.М., Пивкин Е.Н.* Оценка уровня информационной безопасности на основе нечетких моделей с лингвистической шкалой в территориальных налоговых органах // Математическое образование в регионах России: труды Международной научно-практической конференции. – Барнаул: Изд-во АлтГТУ, 2007. – С. 105–108.

5. *Пивкин Е.Н., Белов В.М.* Классификация неоднозначной исходной информации при обеспечении информационной безопасности в территориальных налоговых органах // Управление созданием и развитием систем, сетей и устройств телекоммуникаций: труды научно-практической конференции / под ред. А.В. Бабкина, В.А. Кежаева. – СПб., 2008. – С. 356–362.

6. *Белов В.М., Пивкин Е.Н.* Подход к оценке уровня информационной безопасности в территориальных налоговых органах // Математические методы в технике и технологиях – ММТТ. – 2007. – Т. 10. – С. 239–240.

7. *Капустин А.В., Пивкин Е.Н., Белов В.М.* Оценка уровня информационной безопасности организаций банковской системы на основе аппарата нечетких множеств // Наука и молодежь – 2007: материалы V Всероссийской научно-технической конференции студентов, аспирантов и молодых ученых. – Барнаул: Изд-во АлтГТУ, 2008. – С. 13–15.

8. *Шуроватов М.А. Пивкин Е.Н., Белов В.М.* Оценка уровня защищенности виртуального канала на основе аппарата нечетких множеств // Наука и молодежь – 2007: материалы V Всероссийской научно-технической конференции студентов, аспирантов и молодых ученых. – Барнаул: Изд-во АлтГТУ, 2008. – С. 5–7.

9. *Пивкин Е.Н., Белов В.М.* Нечеткие модели в системе защиты информации, составляющей налоговую тайну // Доклады Томского государственного университета систем управления и радиоэлектроники. – 2008. – № 2 (18), ч. 1. – С. 125–128.

10. *Белов В.М., Пивкин Е.Н., Прокопец В.Д.* Оценка уровня информационной безопасности на основе нечетких моделей в территориальных налоговых органах // Научно-технические ведомости СПбГПУ. – 2007. – № 4-2 (52). – С. 130–132.

11. *Пивкин Е.Н., Белов В.М.* Современные экспертные нечеткие системы: анализ проблемы применения // Снижение рисков и смягчение последствий чрезвычайных ситуаций природного и техногенного характера – приоритетные направления обеспечения комплексной безопасности населения юга Западной Сибири: материалы Шестой международной научно-практической конференции. – Барнаул: Азбука, 2008. – С. 199–200.

12. *Шуроватов М.А., Пивкин Е.Н., Белов В.М.* Программное обеспечение оценки уровня защищенности виртуального канала // Ползуновский альманах. – 2008. – № 4. – С. 90–93.

13. *Капустин А.В., Пивкин Е.Н., Белов В.М.* Программное обеспечение оценки защищенности объектов информатизации на основе аппарата нечетких множеств и стандарта Банка России // Ползуновский альманах. – 2008. – № 4. – С. 97–100.

14. *Шуроватов М.А. Пивкин Е.Н., Белов В.М.* Анализ методов построения функций принадлежности для оценки уровней информационной безопасности на нечетких множествах // Молодежь. Общество. Современная наука, техника и инновации: тезисы докладов VII Всероссийской научной студенческой конференции с международным участием на иностранных языках (8 мая 2008, г. Красноярск) / Сибирский государственный аэрокосмический университет. – Красноярск, 2008. – С. 61–63.

15. *Vopěnka P.* Mathematics in the alternative set theory. – Leipzig: Teubner, 1979. – 120 p.

16. *Vopěnka P.* Fundamentals to mathematics in the alternative set theory. – Bratislava: Alfa, 1990. – 443 p. – In Slovak.

17. *Novák V.* The alternative mathematical model of linguistic semantics and pragmatics. – New York: Plenum, 1992. – 204 p.

Белов Виктор Матвеевич, доктор технических наук, профессор, профессор кафедры защиты информации Новосибирского государственного технического университета. Основное направление научных исследований – применение математических методов в различных областях науки, техники, общества. Имеет более 600 научных публикаций. E-mail: vmbelov@mail.ru.

Пивкин Евгений Николаевич, кандидат технических наук, руководитель направления отдела защиты информации департамента безопасности ПАО АКБ «Связь-Банк». Основное направление научных исследований – применение математических методов в различных областях науки, техники, общества. Автор более 80 публикаций. E-mail: evpiv@yandex.ru.

Ардаева Анастасия Андреевна, магистрант кафедры вычислительной техники Новосибирского государственного технического университета. Область научных интересов – прикладная информатика, информационная безопасность. E-mail: nastya.ardaeva@yandex.ru.

DOI: 10.17212/2782-2230-2022-1-9-26

An integrated approach to assessing the security of significant objects of critical information infrastructure from unauthorized access*

V.M. Belov[1], E.N. Pivkin[2], A.A. Ardaeva[3]

[1] *Novosibirsk State Technical University, Karl Marx Avenue, 20, Novosibirsk, 630073, Russian Federation, doctor of technical sciences, Professor of the department of information security. E-mail: vmbelov@mail.ru*
[2] *PJSC JSCB "Svyaz-Bank", 31/7st. Novoryazanskaya, Moscow, 105066, Russian Federation, candidate of technical sciences, E-mail: evpiv@yandex.ru*
[3] *Novosibirsk State Technical University, Karl Marx Avenue, 20, Novosibirsk, 1630073, Russian Federation, Master's student of the department of computer science. E-mail: nastya.ardaeva@yandex.ru*

The number of cybersecurity events in the modern world has increased significantly, of which a significant number fall on objects of critical information infrastructure. This paper discusses the main requirements for models for assessing the security of significant objects of critical information infrastructure from unauthorized access, classification of ambiguous source information, a generalized algorithm for assessing the level of security of significant objects of critical information infrastructure from unauthorized access, a fuzzy model for assessing the level of security using point and linguistic scales. It also determines the procedure for assessing the security of significant objects of critical information infrastructure from unauthorized access. Among the main requirements for models for assessing the security of significant objects of

* *Received 02 February 2022.*

critical information infrastructure from unauthorized access are the following: versatility, extensibility, formalizability, simplicity, multifactoriality. The generalized algorithm for assessing the level of security of significant objects of critical information infrastructure from unauthorized access includes four computational blocks: in block 1, the collection and primary processing of information is carried out; in block 2, calculations using fuzzy evaluation algorithms with linguistic and point scales are used; in block 3, the effectiveness of information security employees is evaluated; in block 4, the levels of security of significant objects of critical information infrastructure from unauthorized access are predicted using fuzzy time series; in block 5, the estimates obtained at previous stages of calculations are examined using data processing techniques and draw appropriate conclusions for decision-making.

Keywords: security assessment models, classification of initial information, security level assessment algorithm, fuzzy security level assessment model, scoring scale, linguistic scale, security assessment order, unauthorized access

REFERENCES

1. Belov V.M., Pivkin E.N. [Analysis of the problems of creating a model of an integrated information security system in regional tax authorities]. *Aktual'nye problemy bezopasnosti informatsionnykh tekhnologii*: materialy I Mezhdunarodnoi zaochnoi nauchno-prakticheskoi konferentsii [Actual problems of information technology security. Materials of the I International Correspondence Scientific and Practical Conference]. Krasnoyarsk, 2007, pp. 19–21. (In Russian).

2. Belov V.M., Pivkin E.N. Informatsionnye potoki i sposoby ikh formalizatsii s ispol'zovaniem apparata nechetkikh mnozhestv v regional'nykh nalogovykh organakh [Information flows and methods of their formalization using the apparatus of fuzzy sets in regional tax authorities]. *Infokommunikatsionnye sistemy i tekhnologii: problemy i perspektivy* [Infocommunication systems and technologies: problems and prospects]. St. Petersburg, 2007, pp. 336–339.

3. Belov V.M., Pivkin E.N., Prokopets V.D. Osnovnye matematicheskie modeli i metody pri postroenii sistem zashchity informatsii [Basic mathematical models and methods in the construction of information security systems]. *Infokommunikatsionnye sistemy i tekhnologii: problemy i perspektivy* [Infocommunication systems and technologies: problems and prospects]. St. Petersburg, 2007, pp. 300–331.

4. Belov V.M., Pivkin E.N. [Assessment of the level of information security based on fuzzy models with a linguistic scale in the territorial tax authorities]. *Matematicheskoe obrazovanie v regionakh Rossii*: trudy Mezhdunarodnoi nauchno-prakticheskoi konferentsii [Proceedings International scientific-practical conference "Mathematical education in the regions of Russia"]. Barnaul, 2007, pp. 105–108. (In Russian).

5. Pivkin E.N., Belov V.M. [Classification of ambiguous initial information while ensuring information security in the territorial tax authorities]. *Upravlenie sozdaniem i razvitiem sistem, setei i ustroistv telekommunikatsii*: trudy nauchno-

prakticheskoi konferentsii [Management of the creation and development of systems, networks and telecommunications devices. Proceedings of scientific and practical conference]. St. Petersburg, 2008, pp. 356–362. (In Russian).

6. Belov V.M., Pivkin E.N. Podkhod k otsenke urovnya informatsionnoi bezopasnosti v territorial'nykh nalogovykh organakh [Approach to assessing the level of information security in the territorial tax authorities]. *Matematicheskie metody v tekhnike i tekhnologiyakh – MMTT = Mathematical Methods in Technics and Technologies – MMTT*, 2007, vol. 10, pp. 239–240.

7. Kapustin A.V., Pivkin E.N., Belov V.M. [Estimation of the level of information security of organizations of the banking system based on the apparatus of fuzzy sets]. *Nauka i molodezh' – 2007*: materialy V Vserossiiskoi nauchno-tekhnicheskoi konferentsii studentov, aspirantov i molodykh uchenykh [Materials of the V All-Russian scientific and technical conference of students, graduate students and young scientists "Science and youth – 2007"]. Barnaul, 2008, pp. 13–15. (In Russian).

8. Shurovatov M.A. Pivkin E.N., Belov V.M. [Estimation of the security level of a virtual channel based on the apparatus of fuzzy sets]. *Nauka i molodezh' – 2007*: materialy V Vserossiiskoi nauchno-tekhnicheskoi konferentsii studentov, aspirantov i molodykh uchenykh [Materials of the V All-Russian scientific and technical conference of students, graduate students and young scientists "Science and youth – 2007"]. Barnaul, 2008, pp. 5–7. (In Russian).

9. Pivkin E.N., Belov V.M. Nechetkie modeli v sisteme zashchity informatsii, sostavlyayushchei nalogovuyu tainu [Fuzzy models in system of protection of the information containing tax secret]. *Doklady Tomskogo gosudarstvennogo universiteta sistem upravleniya i radioelektroniki = Proceedings of TUSUR University*, 2008, no. 2 (18), pt. 1, pp. 125–128.

10. Belov V.M., Pivkin E.N., Prokopets V.D. Otsenka urovnya informatsionnoi bezopasnosti na osnove nechetkikh modelei v territorial'nykh nalogovykh organakh [Assessment of the level of information security based on fuzzy models in the territorial tax authorities]. *Nauchno-tekhnicheskie vedomosti Sankt-Peterburgskogo gosudarstvennogo politekhnicheskogo universiteta = St. Petersburg State Polytechnical University Journal*, 2007, no. 4-2 (52), pp. 130–132.

11. Pivkin E.N., Belov V.M. [Modern expert fuzzy systems: analysis of the application problem]. *Snizhenie riskov i smyagchenie posledstvii chrezvychainykh situatsii prirodnogo i tekhnogennogo kharaktera – prioritetnye napravleniya obespecheniya kompleksnoi bezopasnosti naseleniya yuga Zapadnoi Sibiri*: materialy Shestoi mezhdunarodnoi nauchno-prakticheskoi konferentsii [Proceedings of the Sixth International scientific-practical conference "Risk reduction and mitigation of the consequences of natural and man-made emergencies – priority areas for

ensuring social security of the population in the South of Western Siberia"]. Barnaul, 2008, pp. 199–200. (In Russian).

12. Shurovatov M.A., Pivkin E.N., Belov V.M. Programmnoe obespechenie otsenki urovnya zashchishchennosti virtual'nogo kanala [Virtual channel security level assessment software]. *Polzunovskii al'manakh = Polzunov Almanac*, 2008, no. 4, pp. 90–93.

13. Kapustin A.V., Pivkin E.N., Belov V.M. Programmnoe obespechenie otsenki zashchishchennosti ob"ektov informatizatsii na osnove apparata nechetkikh mnozhestv i standarta Banka Rossii [Software for assessing the security of informatization objects based on the apparatus of fuzzy sets and the standard of the Bank of Russia]. *Polzunovskii al'manakh = Polzunov Almanac*, 2008, no. 4, pp. 97–100.

14. Shurovatov M.A. Pivkin E.N., Belov V.M. [Analysis of methods for constructing membership functions for assessing information security levels on fuzzy sets]. *Molodezh'. Obshchestvo. Sovremennaya nauka, tekhnika i innovatsii*: tezisy dokladov VII Vserossiiskoi nauchnoi studencheskoi konferentsii s mezhdunarodnym uchastiem na inostrannykh yazykakh [Youth. Society. Modern science, technologies and innovations. Abstracts]. Krasnoyarsk, 2008, pp. 61–63. (In Russian).

15. Vopěnka P. *Mathematics in the alternative set theory*. Leipzig, Teubner, 1979. 120 p.

16. Vopěnka P. *Fundamentals of the mathematics in the alternative set theory*. Bratislava, Alfa, 1990. 443 p. (In Slovak).

17. Novák V. *The alternative mathematical model of linguistic semantics and pragmatics*. New York, Plenum, 1992. 204 p.

Для цитирования:

Белов В.М., Пивкин Е.Н., Ардаева А.А. Комплексный подход к оцениванию защищенности значимых объектов критической информационной инфраструктуры от несанкционированного доступа // Безопасность цифровых технологий. – 2022. – № 1 (104). – С. 9–26. – DOI: 10.17212/2782-2230-2022-1-9-26.

For citation:

Belov V.M., Pivkin E.N., Ardaeva A.A. Kompleksnyi podkhod k otsenivaniyu zashchishchennosti znachimykh ob"ektov kriticheskoi informatsionnoi infrastruktury ot nesanktsionirovannogo dostupa [An integrated approach to assessing the security of significant objects of critical information infrastructure from unauthorized access]. *Bezopasnost' tsifrovykh tekhnologii = Digital Technology Security*, 2022, no. 1 (104), pp. 9–26. DOI: 10.17212/2782-2230-2022-1-9-26.

БЕЗОПАСНОСТЬ ЦИФРОВЫХ ТЕХНОЛОГИЙ. – 2022. – № 1 (104). – 27–40

МЕТОДЫ И СИСТЕМЫ ЗАЩИТЫ ИНФОРМАЦИИ, ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ

УДК 004.9 DOI: 10.17212/2782-2230-2022-1-27-40

АНАЛИЗ МЕТОДИЧЕСКОГО И ТЕХНИЧЕСКОГО ОБЕСПЕЧЕНИЯ ПРОЦЕДУР ОЦЕНИВАНИЯ ЗАЩИЩЕННОСТИ ЗНАЧИМЫХ ОБЪЕКТОВ КРИТИЧЕСКОЙ ИНФОРМАЦИОННОЙ ИНФРАСТРУКТУРЫ ОТ НЕСАНКЦИОНИРОВАННОГО ДОСТУПА*

Е.Н. ПИВКИН[1], А.А. АРДАЕВА[2]

[1] *105066, РФ, г. Москва, ул. Новорязанская, 31/7, к. 2, ПАО АКБ «Связь-Банк», кандидат технических наук. E-mail: evpiv@yandex.ru*
[2] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, магистрант кафедры вычислительной техники. E-mail: nastya.ardaeva@yandex.ru*

В настоящей работе проводится непосредственный анализ существующих методик оценивания защищенности значимых объектов критической информационной инфраструктуры, обзор их законодательных основ и существующих средств защиты информации от несанкционированного доступа. Такой анализ необходим для решения задач, связанных с разработкой комплексного подхода к оцениванию защищенности значимых объектов критической информационной инфраструктуры. Рассмотрены основные руководящие документы и приказы ФСТЭК России, Федеральный закон № 187-ФЗ от 26 июля 2017 г. «О безопасности КИИ РФ». Проанализирован современный рынок средств защиты информации от несанкционированного доступа. Для удобства все сравнительные критерии разделены на категории: общие сведения; системные требования (минимальные); поддерживаемые автоиатизированные рабочие места и серверы на основе известных защищенных операционных систем; уровень сертификации по требованиям безопасности ФСТЭК России; развертывание системы защиты; обновление компонентов; основные функции средств защиты информации от несанкционированного доступа; очистка информации; дополнительные модули защиты; централизованное управление и отчетность; возможность интеграции; лицензирование. Для участия в сравнении были выбраны четыре наиболее популярные в России группы средств защиты информации от несанкционированного доступа: Secret Net Studio; Dallas Lock 8.0-К; Diamond ACS; «Блокхост-Сеть 2.0». С целью выявления методик оценивания защищенности значимых объектов критической информационной инфраструктуры рассмотрены национальные стандарты России и

* Статья получена 04 февраля 2022 г.

научная периодика. Показано, что методическое обеспечение данного сегмента безопасности находится не на должном уровне.

Ключевые слова: сравнительный анализ, методики оценивания защищенности, значимые объекты, критическая информационная инфраструктура, информационная безопасность, законодательные основы, средства защиты информации, несанкционированный доступ

ВВЕДЕНИЕ

После вступления в силу Федерального закона от 26.07.2017 № 187-ФЗ «О безопасности критической информационной инфраструктуры Российской Федерации» (КИИ) [1], регулирующего деятельность по обеспечению безопасности объектов информационной инфраструктуры РФ, функционирование которых критически важно для экономики государства, актуальность построения систем защиты информации (СЗИ) на объектах КИИ не вызывает сомнений. Документ определяет описание субъекта и объекта КИИ, безопасности КИИ и компьютерного инцидента. В Постановлении Правительства РФ от 08.02.2018 № 127 «Об утверждении Правил категорирования объектов критической информационной инфраструктуры Российской Федерации, а также перечня показателей критериев значимости объектов критической информационной инфраструктуры Российской Федерации и их значений» [2] определены правила категорирования объектов КИИ. Также принят Приказ ФСТЭК № 239 «Об утверждении Требований по обеспечению безопасности значимых объектов критической информационной инфраструктуры РФ» от 25.12.2017 [3]. Деятельность, относящаяся к обеспечению информационной безопасности (ИБ), должна быть качественно организована и проконтролирована. Для этого необходимо регулярное оценивание уровня защищенности значимых объектов критической информационной инфраструктуры (ЗО КИИ), рисков и мер, принимаемых для управления этими рисками.

Целью настоящей работы является сравнительный анализ методик оценивания защищенности ЗО КИИ от несанкционированного доступа (НСД), их законодательной основы и существующих СЗИ от НСД.

1. АНАЛИЗ НОРМАТИВНО-ПРАВОВЫХ ДОКУМЕНТОВ ПО ОЦЕНИВАНИЮ ЗАЩИЩЕННОСТИ ЗО КИИ

В этом разделе рассмотрим наиболее значимые руководящие документы ФСТЭК России, устанавливающие критерии для оценивания защищенности объектов информатизации.

В Руководящем документе «СВТ. Защита от НСД к информации. Показатели защищенности от НСД к информации» [4] приводится классификация средств по уровню безопасности от НСД на основе перечня показателей защищенности и совокупности описывающих их требований. Главной основой для разработки данного документа послужила «Оранжевая книга». Этот стандарт устанавливает семь классов защищенности СВТ (средства вычислительной техники) от НСД к информации.

Классы делятся на четыре группы, которые различаются по уровню защиты: в первую группу входит лишь седьмой класс, к которому относят все средства, не соответствующие требованиям более высоких классов; во вторую группу входят шестой и пятый классы, и она характеризуется дискреционной защитой; в третью группу входят четвертый, третий и второй классы, и она характеризуется мандатной защитой; четвертая группа включает в себя лишь первый класс и характеризуется верифицированной защитой. Самый высокий класс – первый, самый низкий – седьмой.

В Руководящем документе «Автоматизированные системы. Защита от НСД к информации [5]. Классификация автоматизированных систем (АС) и требования по защите информации» приведены классификация АС, подлежащих защите от НСД к информации, и требования по защите информации в АС различных классов.

В документе выделяют девять классов защиты АС от НСД к информации. Каждый класс характеризуется определенным минимальным набором требований к безопасности. Классы делятся на три группы, которые различаются по свойствам обработки информации в АС.

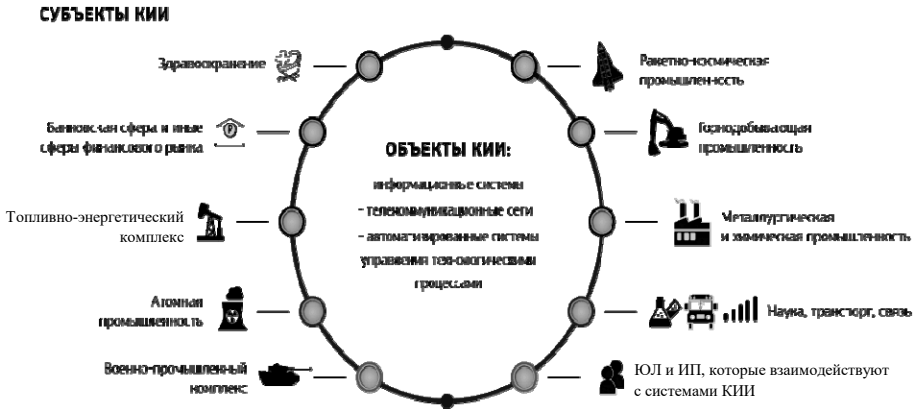
Внутри каждой группы поддерживается иерархия требований безопасности, зависящая от ценности и конфиденциальности информации, и соответственно, иерархия классов защищенности АС.

Руководящий документ «Средства вычислительной техники. Межсетевые экраны. Защита от НСД к информации. Показатели защищенности от НСД к информации» [6] является основным для анализа внешней системы защиты периметра корпоративной сети. Этот документ определяет показатели защищенности межсетевых экранов. Каждый критерий защищенности представляет собой набор требований безопасности, характеризующих конкретную область эксплуатации МЭ. В документе описаны пять показателей защищенности.

В данном случае рассматриваются информационные системы (ИС) ЗО КИИ Российской Федерации. При нарушении функционирования предприятия, которое может вызвать сбой ИС, масштаб ущерба крайне сложно определить.

Базовым законом систем ЗО КИИ является Федеральный закон от 26 июля 2017 г. № 187-ФЗ «О безопасности КИИ РФ», регулирующий деятельность по обеспечению безопасности объектов информационной инфраструктуры РФ, функционирование которых критически важно для экономики государства. Такие объекты в ФЗ называются объектами КИИ.

Федеральный закон № 187 четко определяет, что «к субъектам критической информационной инфраструктуры относятся государственные органы и учреждения, а также российские юридические лица и/или индивидуальные предприниматели, которым на праве собственности, аренды или на ином законном основании принадлежат информационные системы, информационно-телекоммуникационные сети, автоматизированные системы управления» [1]. У каждого субъекта КИИ есть объекты КИИ: ИС, АС управления технологическими процессами, информационно-телекоммуникационные сети (рисунок).



Субъекты и объекты критической информационной инфраструктуры

Subjects and objects of critical information infrastructure

В ФЗ № 187 четко регламентированы обязанности субъекта КИИ: категорирование объектов КИИ; выполнение требований по безопасности ЗО КИИ; обеспечение взаимодействия с государственной системой обнаружения, предупреждения и ликвидации последствий компьютерных атак (ГосСОПКА). При рассмотрении систем ЗО КИИ необходимо учитывать не только ФЗ № 187, но и Приказ ФСТЭК России от 25 декабря 2017 г. № 239, в котором обозначены требования по обеспечению безопасности на всех стадиях

жизненного цикла, а также требования по внедрению организационных и технических мер по обеспечению безопасности ЗО КИИ. В 2019 г. были внесены поправки в приказы ФСТЭК России № 239 [3] и № 17 [7], что ознаменовало переход к новой системе требований, предъявляемых к средствам защиты информации.

В настоящее время система требований к оцениванию основывается на уровнях доверия, устанавливаемых приказом ФСТЭК России № 131 [8]. Дополнительно к нему вышла методика выявления уязвимостей и недекларированных возможностей в программном обеспечении. Выполнение требований к уровню доверия является обязательным при проведении работ по оцениванию соответствия, в том числе и в форме сертификации СЗИ, организуемых ФСТЭК России в пределах своих полномочий. Требования к разработке и производству СЗИ, к проведению испытаний СЗИ, а также к поддержке безопасности СЗИ в ходе их применения устанавливаются требованиями к уровням доверия. Для дифференциации указанных требований устанавливается шесть уровней доверия. Самый низкий уровень – шестой, самый высокий – первый [9].

2. АНАЛИЗ СРЕДСТВ ЗАЩИТЫ ОТ НСД ДЛЯ ЗО КИИ

Для удобства анализа современных СЗИ все сравнительные критерии были разделены на следующие категории: общие сведения; системные требования (минимальные); поддерживаемые автоматизированные рабочие места (АРМ) и серверы на основе известных защищенных операционных систем; уровень сертификации по требованиям безопасности ФСТЭК России; развертывание системы защиты; обновление компонентов; основные функции СЗИ от НСД; очистка информации; дополнительные модули защиты; централизованное управление и отчетность; возможность интеграции; лицензирование.

Для участия в сравнении были выбраны четыре наиболее популярные в России СЗИ от НСД: Secret Net Studio; Dallas Lock 8.0-К; Diamond ACS; «Блокхост-Сеть 2.0».

Исходя из анализа средств защиты информации от несанкционированного доступа, можно сделать вывод о том, что комплекс хорошо описанных организационных процессов и общей архитектуры системы защиты имеет гораздо большее значение, нежели отдельные внедряемые механизмы безопасности (таблица).

Сравнение СЗИ от НСД

Comparison of SPI from NSD

Параметр сравнения	Secret Net Studio	Dallas Lock	Diamond ACS	Блокхост-Сеть
Компания-вендор	ООО «Код Безопасности»	ООО «Конфидент»	ООО «ТСС»	ООО «Газинформсервис
Целевой сегмент	Крупный бизнес, средний бизнес. Государственный сектор			
Интерфейс	Русский, английский	Русский	Русский, английский	Русский, английский
Требования к серверу безопасности системы	CPU Intel Core i5/Intel Xeon E3, RAM 8 GB, HDD 150 GB	Конфигурация сервера определяется требованиями к соответствующей ОС. Для размещения файлов СЗИ НСД требуется не менее 200 MB HDD	Конфигурация сервера определяется требованиями к соответствующей ОС	Конфигурация сервера определяется требованиями к соответствующей ОС из списка разрешенных к применению
Поддерживаемые операционные системы	Windows Server 2008 R2, Windows Server 2012/2012R2, Windows Server 2016, windows Vista, Windows 7, Windows 8/8.1, Windows 10			
Уровень сертификации по требованиям безопасности ФСТЭК России	РД СВТ – 5-й класс защищенности СВТ от НСД	РД СВТ – 5-й класс защищенности СВТ от НСД	РД СВТ – 3-й класс защищенности СВТ от НСД	РД СВТ – 5-й класс защищенности СВТ от НСД

Продолжение таблицы

Table continuation

Параметр сравнения	Secret Net Studio	Dallas Lock	Diamond ACS	Блокхост-Сеть
	РД НДВ – 4-й уровень контроля отсутствия НДВ МЭ – 4-й класс защиты МЭ СКН – 4-й класс защиты СКН САВЗ – 4-й класс защиты САВЗ СОВ – 4-й класс защиты СОВ	РД НДВ – 4-й уровень контроля отсутствия НДВ МЭ – нет СКН – 4-й класс защиты СКН САВЗ – нет СОВ – 4-й класс защиты СОВ	РД НДВ – 2-й уровень контроля отсутствия НДВ МЭ – нет СКН – нет САВЗ – нет СОВ – нет	РД НДВ – 4-й уровень контроля отсутствия НДВ МЭ – класс защиты МЭ СКН – 4-й класс защиты СКН САВЗ – нет СОВ – нет
Параметр сравнения	Secret Net Studio	Dallas Lock	Diamond ACS	Блокхост-Сеть
Контроль входа в систему, блокировка сеанса работы	Да	Да	Да	Да
Контроль объектов системы	Да	Да	Да	Да
Контроль целостности системы	Да	Да	Нет	Нет
Контроль портов ввода/вывода подключаемых устройств	Да	Да	Да	Да
Персональный межсетевой экран	Да	Да	Нет	Да

Окончание таблицы

End of Table

Параметр сравнения	Secret Net Studio	Dallas Lock	Diamond ACS	Блокхост-Сеть
Средство обнаружения вторжения уровня хоста	Да	Да	Нет	Нет
Средство антивирусной защиты	Да	Нет	Нет	Нет
Управление и мониторинг СЗИ	Да	Да	Да (для автономной версии)	Да
Регистрация и учет событий	Да	Да	Да	Да
Описание политики лицензирования	По числу защищаемых компьютеров / по защитным модулям	По числу защищаемых компьютеров / по защитным модулям	По числу рабочих мест. Лицензия на использование сервера управления приобретается отдельно	По числу рабочих мест. Лицензия на использование сервера управления приобретается отдельно

3. АНАЛИЗ СУЩЕСТВУЮЩИХ МЕТОДИК ОЦЕНИВАНИЯ ЗАЩИЩЕННОСТИ ОТ НСД ДЛЯ ЗО КИИ

Разработку методических рекомендаций по оцениванию защищенности объектов информатизации и изменению структуры системы ее защиты проводили, например, Л.Г. Осовецкий и С.В. Максименко применительно к АС обработки данных, однако систему в целом для ЗО КИИ не рассматривали, для оценивания вероятности конкретной угрозы использовали экспертные оценки.

В Приказе ФСТЭК России № 131 отражены в основном требования к защите ЗО КИИ, но пользователям не предоставляются ни методология оцени-

вания эффективности, ни требования к разработке, оформлению и содержанию необходимых документов.

ГОСТ Р ИСО/МЭК ТО 19791–2008 «Информационная технология. Методы и средства обеспечения безопасности. Оценка безопасности автоматизированных систем» [10] содержит рекомендации и критерии оценивания безопасности АС, а также обеспечивает расширение области применения стандартов серии ИСО/МЭК 15408, включая ряд критических аспектов, касающихся оценивания среды эксплуатации объекта, оценивания и декомпозиции составных АС на домены безопасности, которые должны оцениваться отдельно. Устанавливает определение и модель АС; описывает расширения концепции оценивания безопасности с помощью стандартов серии ИСО/МЭК 15408; методологию и процесс выполнения оценивания безопасности АС; дополнительные критерии оценивания безопасности, охватывающие те аспекты АС, которые не были охвачены критериями оценивания безопасности в стандартах серии ИСО/МЭК 15408. Дает возможность включать продукты безопасности, оцененные в соответствии с требованиями стандартов серии ИСО/МЭК 15408, в автоматизированные системы и проводить оценку как единого целого с использованием настоящего стандарта.

Исходя из анализа национального стандарта РФ ГОСТ Р ИСО/МЭК 15408-1–2012 [11], примерами объекта оценивания являются прикладная программа; операционная система; прикладная программа в сочетании с операционной системой; прикладная программа в сочетании с операционной системой и рабочей станцией; операционная система в сочетании с рабочей станцией; интегральная схема смарт-карты; локальная вычислительная сеть, включая все терминалы, серверы, сетевое оборудование и программные средства; приложение базы данных, за исключением программных средств удаленного клиента, обычно ассоциируемых с приложением базы данных.

В публикациях [12–25] приведены алгоритмы и модели оценивания защищенности некоторых объектов территориальных налоговых органов (ТНО) Российской Федерации (РФ). Авторы приведенных работ не сформулировали методики оценивания защищенности ТНО РФ, это в особенности касается ЗО КИИ, проблематика которых возникла в результате вступления в силу Федерального закона № 187. С учетом существующих разработок для ТНО РФ работы [12–25], на наш взгляд, представляют интерес для дальнейшего развития в рамках модификации и апробации для ЗО КИИ, учитывая, что методические разработки в этом направлении отсутствуют.

ЗАКЛЮЧЕНИЕ

Таким образом, федеральные органы исполнительной власти, которые отвечают за обеспечение ИБ КИИ, действуют на основании своих нормативных документов, в которых отсутствуют описание единого регламента проведения аудита ЗО КИИ и оценочные показатели соответствия требованиям ИБ.

В области ИБ нет национального стандарта, устанавливающего единые регламенты и процедуры проведения аудита, тем более для аудита КИИ (методики, описывающие порядок действий по определению уровня защищенности ОИ, показатели ИБ, способы оценивания показателей и уровней защищенности), общей методики проведения оценивания защищенности объектов информатизации, в том числе и ЗО КИИ, автоматизированных методик даже в рамках экспертного подхода.

БЛАГОДАРНОСТИ

Автор выражает глубокую благодарность д-ру техн. наук, профессору Белову Виктору Матвеевичу за ценные советы и замечания, высказанные при работе над статьей.

СПИСОК ЛИТЕРАТУРЫ

1. Федеральный закон от 26.07.2017 № 187-ФЗ «О безопасности критической информационной инфраструктуры Российской Федерации».

2. Постановление Правительства Российской Федерации «Об утверждении правил категорирования объектов критической информационной инфраструктуры Российской Федерации, а также перечня показателей критериев значимости объектов критической информационной инфраструктуры Российской Федерации и их значений» от 08.02.2018 № 127 (с изм. и доп. в ред. от 13 апреля 2019 г.).

3. Приказ ФСТЭК России от 25.12.2017 № 239 «Об утверждении требований по обеспечению безопасности значимых объектов критической информационной инфраструктуры Российской Федерации» (в ред. Приказов ФСТЭК России от 9 августа 2018 г. № 138, от 26 марта 2019 г. № 60, от 20 февраля 2020 г. № 35).

4. Руководящий документ «Средства вычислительной техники. Защита от несанкционированного доступа к информации. Показатели защищенности от несанкционированного доступа к информации»: утв. решением Государственной технической комиссии при Президенте Российской Федерации от 30 марта 1992 г.

5. Руководящий документ «Автоматизированные системы. Защита от НСД к информации. Классификация автоматизированных систем и требования по защите информации»: утв. решением Государственной технической комиссии при Президенте Российской Федерации от 30 марта 1992 г.

6. Приказ ФСТЭК России от 11.02.2013 № 17 «Об утверждении требований о защите информации, не составляющей государственную тайну, содержащейся в государственных информационных системах».

7. Приказ ФСТЭК России от 21.09.2016 № 131 «Об организации работы по разработке перечней правовых актов и их отдельных частей (положений), содержащих обязательные требования, соблюдение которых оценивается при проведении мероприятий по контролю в рамках видов федерального государственного контроля в сфере компетенции ФСТЭК России».

8. Информационное сообщение «О требованиях безопасности информации, устанавливающих уровни доверия к средствам технической защиты информации и средствам обеспечения безопасности информационных технологий» от 29.03.2019 № 240/24/1525.

9. ГОСТ Р ИСО/МЭК ТО 19791–2008. Информационная технология. Методы и средства обеспечения безопасности. Оценка безопасности автоматизированных систем: введ. 2009–10–01. – М.: Стандартинформ, 2010. – 121 с.

10. ГОСТ Р ИСО/МЭК 15408-1–2012. Информационная технология. Методы и средства обеспечения безопасности. Критерии оценки безопасности информационных технологий. Ч. 1. Введение и общая модель: взамен ГОСТ Р ИСО/МЭК 15408-1–2008: введ. 2013–12–01. – М.: Стандартинформ, 2010. – 51 с.

11. *Пивкин Е.Н., Белов В.М.* Нечеткие модели в системе защиты информации, составляющей налоговую тайну // Доклады Томского государственного университета систем управления и радиоэлектроники. – 2008. – № 2 (18), ч. 1. – С. 125–128.

12. *Белов В.М., Пивкин Е.Н., Прокопец В.Д.* Оценка уровня информационной безопасности на основе нечетких моделей в территориальных налоговых органах // Научно-технические ведомости СПбГПУ. – 2007. – № 4-2 (52). – С. 130–132.

13. *Шуроватов М.А., Пивкин Е.Н., Белов В.М.* Программное обеспечение оценки уровня защищенности виртуального канала // Ползуновский альманах. – 2008. – № 4. – С. 90–93.

14. *Капустин А.В., Пивкин Е.Н., Белов В.М.* Программное обеспечение оценки защищенности объектов информатизации на основе аппарата нечетких множеств и стандарта Банка России // Ползуновский альманах. – 2008. – № 4. – С. 97–100.

Пивкин Евгений Николаевич, кандидат технических наук, руководитель направления отдела защиты информации департамента безопасности ПАО АКБ «Связь-Банк». Основное направление научных исследований – применение математических методов в различных областях науки, техники, общества. Автор более 80 публикаций. E-mail*:* evpiv@yandex.ru

Ардаева Анастасия Андреевна, магистрант кафедры вычислительной техники Новосибирского государственного технического университета. Область научных интересов – прикладная информатика, информационная безопасность. E-mail: nastya.ardaeva@yandex.ru

DOI: 10.17212/2782-2230-2022-1-27-40

Analysis of methodological and technical support of procedures for assessing the security of significant objects of critical information infrastructure from unauthorized access*

E.N. Pivkin[1], A.A. Ardaeva[2]

[1] *PJSC JSCB "Svyaz-Bank", 31/7 Street Novoryazanskaya, Moscow, 105066, Russian Federation, candidate of technical sciences, E-mail: evpiv@yandex.ru*
[2] *Novosibirsk State Technical University, Karl Marx Avenue, 20, Novosibirsk, 1630073, Russian Federation, Master's student of the Department of Computer Science. E-mail: nastya.ardaeva@yandex.ru*

This paper provides a direct analysis of existing methods for assessing the security of significant objects of critical information infrastructure, a review of their legislative framework and existing means of protecting information from unauthorized access. Such an analysis is necessary to solve problems related to the development of an integrated approach to assessing the security of significant objects of critical information infrastructure. The main guiding documents and orders of the FSTEC of Russia, Federal Law No. 187-FZ of July 26, 2017 "On the security of the CII of the Russian Federation" were considered. The modern market of means of protecting information from unauthorized access was analyzed. For convenience, all comparative criteria were divided into categories: general information; system requirements (minimum); supported automated workstations and servers based on well-known secure operating systems; the level of certification according to the safety requirements of the FSTEC of Russia; deployment of a protection system; component updates; the main functions of the means of protecting information from unauthorized access; clearing information; additional protection modules; centralized management and reporting; possibility of integration; licensing. The four most popular Russian groups of means of protecting information from unauthorized access were selected to participate in the comparison: Secret Net Studio; Dallas Lock 8.0-K; Diamond ACS; Blockhost Network 2.0. In order to identify methods for assessing the security of significant objects of critical information infrastructure, national standards of Russia and

* *Received 04 February 2022.*

scientific periodicals were considered. It is shown that the methodological support of this segment of safety is not at the proper level.

Keywords: comparative analysis, security assessment methods, significant objects, critical information infrastructure, information security, legal framework, information security tools, unauthorized access

REFERENCES

1. RF Federal Law “On the security of the critical information infrastructure of the Russian Federation” of July 26, 2017 N 187. (In Russian).

2. Decree of the Government of Russia N 127 of February 8, 2018 “On approval of the rules for categorizing the objects of critical information infrastructure of the Russian Federation, and also the list of indicators of the criteria of significance of the objects of critical information infrastructure of the Russian Federation and their values” (as amended on April 13, 2019). (In Russian).

3. Order of the FSTEC of Russia of December 25, 2017 N 239 “On approval of the requirements for ensuring the security of significant objects of critical information infrastructure of the Russian Federation” (as amended by the Orders of the FSTEC of Russia of August 9, 2018 N 138, of March 26, 2019 N 60, of February 20, 2020 N 35). (In Russian).

4. Guidance Document “Computer facilities. Protection against unauthorized access to information. Indicators of security against unauthorized access to information”. Approved by the decision of the State Technical Commission under the President of the Russian Federation of March 30, 1992. (In Russian).

5. Guidance Document “Automated systems. Protection from unauthorized access to information. Classification of automated systems and information security requirements”. Approved by the decision of the State Technical Commission under the President of the Russian Federation of March 30, 1992. (In Russian).

6. Order of the FSTEC of Russia of February 11, 2013 N 17 “On approval of requirements for the protection of information not constituting a state secret contained in state information systems”. (In Russian).

7. Order of the FSTEC of Russia of September 21, 2016 N 131 “On the organization of work on the development of lists of legal acts and their separate parts (provisions) containing mandatory requirements, compliance with which is assessed when carrying out control measures within the framework of types of federal state control within the competence of the FSTEC of Russia”. (In Russian).

8. Information message “On information security requirements that establish levels of trust in technical information security tools and information technology security tools” of March 29, 2019 N 240/24/1525. (In Russian).

9. State Standard R ISO/IEC TO 19791–2008. *Information technology. Security techniques. Security assessment of operational systems*. Moscow, Standartinform Publ., 2010. 121 p. (In Russian).

10. State Standard R ISO/IEC 15408-1–2012. *Information technology. Security techniques. Evaluation criteria for IT security*. Pt. 1. *Introduction and general model*. Moscow, Standartinform Publ., 2010. 51 p. (In Russian).

11. Pivkin E.N., Belov V.M. Nechetkie modeli v sisteme zashchity informatsii, sostavlyayushchei nalogovuyu tainu [Fuzzy models in system of protection of the information containing tax secret]. *Doklady Tomskogo gosudarstvennogo universiteta sistem upravleniya i radioelektroniki = Proceedings of TUSUR University*, 2008, no. 2 (18), pt. 1, pp. 125–128.

12. Belov V.M., Pivkin E.N., Prokopets V.D. Otsenka urovnya informatsionnoi bezopasnosti na osnove nechetkikh modelei v territorial'nykh nalogovykh organakh [Assessment of the level of information security based on fuzzy models in the territorial tax authorities]. *Nauchno-tekhnicheskie vedomosti Sankt-Peterburgskogo gosudarstvennogo politekhnicheskogo universiteta = St. Petersburg State Polytechnical University Journal*, 2007, no. 4-2 (52), pp. 130–132.

13. Shurovatov M.A., Pivkin E.N., Belov V.M. Programmnoe obespechenie otsenki urovnya zashchishchennosti virtual'nogo kanala [Virtual channel security level assessment software]. *Polzunovskii al'manakh = Polzunov Almanac*, 2008, no. 4, pp. 90–93.

14. Kapustin A.V., Pivkin E.N., Belov V.M. Programmnoe obespechenie otsenki zashchishchennosti ob"ektov informatizatsii na osnove apparata nechetkikh mnozhestv i standarta Banka Rossii [Software for assessing the security of informatization objects based on the apparatus of fuzzy sets and the standard of the Bank of Russia]. *Polzunovskii al'manakh = Polzunov Almanac*, 2008, no. 4, pp. 97–100.

Для цитирования:

Пивкин Е.Н., Ардаева А.А. Анализ методического и технического обеспечения процедур оценивания защищенности значимых объектов критической информационной инфраструктуры от несанкционированного доступа // Безопасность цифровых технологий. – 2022. – № 1 (104). – С. 27–40. – DOI: 10.17212/2782-2230-2022-1-27-40.

For citation:

Pivkin E.N., Ardaeva A.A. Analiz metodicheskogo i tekhnicheskogo obespecheniya protsedur otsenivaniya zashchishchennosti znachimykh ob"ektov kriticheskoi informatsionnoi infrastruktury ot nesanktsionirovannogo dostupa [Analysis of methodological and technical support of procedures for assessing the security of significant objects of critical information infrastructure from unauthorized access]. *Bezopasnost' tsifrovykh tekhnologii = Digital Technology Security*, 2022, no. 1 (104), pp. 27–40. DOI: 10.17212/2782-2230-2022-1-27-40.

БЕЗОПАСНОСТЬ ЦИФРОВЫХ ТЕХНОЛОГИЙ. – 2022. – № 1 (104). – 41–60

МЕТОДЫ И СИСТЕМЫ ЗАЩИТЫ ИНФОРМАЦИИ, ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ

УДК 004.9 DOI: 10.17212/2782-2230-2022-1-41-60

АЛГОРИТМЫ И МЕТОДЫ КЛАСТЕРИЗАЦИИ ДАННЫХ В АНАЛИЗЕ ЖУРНАЛОВ СОБЫТИЙ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ *

Д.Н. СИДОРОВА[1], Е.Н. ПИВКИН[2]

[1] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, ассистент кафедры защиты информации. E-mail: d.sidorova.2013@corp.nstu.ru*
[2] *105066, РФ, г. Москва, ул. Новорязанская, 31/7, к. 2, ПАО АКБ «Связь-Банк», кандидат технических наук. E-mail: evpiv@yandex.ru*

Файлы журналов регистрации событий безопасности дают представление о состоянии инфосистем и возможности находить аномалии в поведении пользователей, а также диагностировать происшествия кибербезопасности. В работе рассмотрены существующие журналы событий (журналы событий приложений, систем, безопасности). Следует отметить, что автоматический анализ данных журналов событий осложнен, так как они содержат большое количество неструктурированных данных, которые собираются из разных источников. Поэтому в настоящей статье представлена и описана проблема анализа журналов событий информационной безопасности. Для решения проблемы анализа журналов безопасности были рассмотрены новые и не особо изученные методы и алгоритмы кластеризации данных, как Randomforest («случайный лес»), инкрементальная кластеризация, алгоритм Iterative Partitioning Log Mining (IPLoM) – итеративный анализ журналов секционирования. Алгоритм Randomforest создает деревья решений для выборок данных, после чего делается прогноз по каждой выборке и с помощью голосования выбирается наилучшее решение. Такой метод сокращает переобучение путем усреднения показателей. Также алгоритм применяется в таких типах задач, как регрессия и классификация. Инкрементальная кластеризация определяет кластеры как группы объектов, которые принадлежат одному классу или концепту. Когда кластеры определяются, то они могут перекрываться, поэтому допускается степень «размытости для выборок», которые лежат на границах разных кластеров. Алгоритм итеративного анализа журналов секционирования использует уникальные характеристики сообщений журналов для их итеративного разделения, что способствует эффективному извлечению типов сообщений.

Ключевые слова: алгоритмы, методы, кластеризация данных, информационная безопасность, «случайный лес», инкрементальная кластеризация, итеративный анализ журнала секционирования, журналы событий

* Статья получена 07 февраля 2022 г.

ВВЕДЕНИЕ

В научной литературе на данном этапе отсутствует устоявшееся определение понятия «кластеризация данных (кластерный анализ)». В настоящей работе будем использовать определение из словаря [1]: кластерный анализ – совокупность математических методов, предназначенных для формирования относительно «отдаленных» друг от друга групп «близких» между собой объектов по информации о расстояниях или связях (мерах близости) между ними. Кластеризация используется в различных областях деятельности человека, и в каждой индивидуальной задаче ее использование имеет свои особенности. На сегодняшний день существует множество различных методов и алгоритмов кластеризации данных, которые наиболее распространены и используемы [2–14]. Поэтому нашей задачей в настоящей статье будем считать рассмотрение наименее известных алгоритмов и методов кластеризации данных.

1. ЖУРНАЛЫ СОБЫТИЙ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

Событием называется любое существенное изменение состояния системы либо программы, о котором следует уведомить пользователей, также событием называется запись в журнале событий. Служба журнала событий записывает события приложений, системные события и события безопасности в средство просмотра событий. При помощи журналов – инструмента просмотра событий – возможно принимать информацию об аппаратном и программном обеспечении компьютера, о системных компонентах, а также проверять события безопасности на локальных или удаленных компьютерах. Журналы событий допускается применять при распознавании источников текущих системных проблем и для недопущения вероятных проблем [15].

Принято подразделять журналы событий на три вида: журналы событий и приложений, журналы систем, журналы безопасности.

Журналы событий приложений включают события, которые сформированы приложениями, а не системой. Сервер базы данных записывает ошибки, которые происходят при его работе в журнале приложений. Есть возможность для разработчиков самостоятельно принимать решения, какие события оставлять в протоколах журналов событий приложений. К примеру, Майкрософт SQL Server протоколирует детализированные данные о принципиальных аварийных ситуациях, которые возникают при работе SQL-сервера, таких как «недостаточно памяти», «сбой при запасном копировании базы данных» и т. д. При всем этом события, сгенерированные различными приложениями, попадают в единый журнал приложений. Приложения распознаются как различные «источники» в первоначальном свойстве событий. Поэтому нетрудно

выделить события определенного приложения. Коды событий (ID) распознаются приложением, которое сформировало эти коды. События могут повторяться для разных источников.

Журналы событий систем включают события, которые сгенерированы системными элементами. Например, отказы драйверов либо остальных системных компонентов при запуске систем фиксируются в системных журналах событий. Типы и коды событий системных компонентов заранее назначены разработчиками операционных систем (например, Windows). Подобно журналу приложений, системный журнал включает события из различных источников. Стоит обратить внимание, что определенные события идентифицируются и кодом, и источником. Журналы событий систем являются важным информационным источником при поиске обстоятельств отказов и вопросов системных администраторов и технических экспертов [16, 17].

Журналы событий безопасности содержит события, которые оказывают влияние на безопасность систем. Это попытки (удачные и не удачные) входа в профили систем, внедрение ресурсов (файлов, списка, устройств), управление учетными записями, изменения прав и преимуществ учетных записей, пуск и остановка процессов (программ) и т. д. Администратор может изменять категории событий, необходимых для регистрации. По умолчанию система всегда сконфигурирована так, чтобы регистрировать события управления учетными записями, события входа в систему. Как правило, аудит доступа к объектам не включается в записи журналов. В данном случае важно быть осторожным при настройке аудита доступа к файлам: некомпетентная настройка может привести к возникновению значительного числа событий, что, в свою очередь, плохо отразится на общей продуктивности системы и может привести к скорому переполнению журнала безопасности.

Запись в журнал безопасности делается лишь системными элементами, коды событий совершенно точно идентифицируют события. Журналы событий безопасности являются значительным информационным источником при расследовании происшествий безопасности, и их изучение существенно для администраторов безопасности, специалистов по информационной безопасности и специалистов по цифровой криминалистической экспертизе.

Журнал регистрации событий безопасности – это электронный журнал, который включает записи о ситуации кибербезопасности, о действиях пользователей и эксплуатирующего персонала автоматической системы [18].

Журналы событий можно разделить на следующие типы.

1. Журналы, отслеживающие вход / выход пользователя:
   а) журнал регистрации неудачных попыток начала сеанса;
   б) журнал регистрации начала / завершения сеанса.

2. Журналы, отслеживающие события:
   а) журнал регистрации событий;
   б) архив журнала регистрации событий.
3. Журнал регистрации сообщений об ошибках.

2. АЛГОРИТМЫ И МЕТОДЫ КЛАСТЕРИЗАЦИИ

Существует множество различных методов и алгоритмов кластеризации данных. В основном методы кластеризации классифицируют на иерархические и неиерархические (основанные на плоском разделении). Иерархические методы разделяют на агломеративные и дивизимные, неиерархические методы – на четкие и нечеткие. К наиболее популярным алгоритмам и методам кластеризации можно отнести следующие: наивный байесовский подход, методы опорных векторов, ближайших соседей, деревьев решений, искусственных нейронных сетей, нечеткой логики, генетические алгоритмы [19–28].

Поэтому мы рассмотрим наименее известные методы и алгоритмы кластерного анализа, такие как Randomforest («случайный лес»), инкрементальную кластеризацию, алгоритм IPLoM (Iterative Partitioning Log Mining), так как они применимы для больших объемов данных и характеризуются высокой точностью.

Randomforest

«Случайный лес» – это популярная процедура машинного обучения, которую можно использовать для разработки моделей прогнозирования. Впервые представленные Брейманом в 2001 году «случайные леса» представляют собой простые модели из наборов деревьев классификации и регрессии, использующие двоичное разбиение переменных-предикторов для определения прогнозов результатов [29–31].

Деревья решений просты в использовании на практике, так как представляют собой интуитивно понятный метод прогнозирования результата, который разделяет «высокие» и «низкие» значения предиктора, связанного с результатом. Несмотря на то что данный метод предлагает множество преимуществ, методология дерева решений часто обеспечивает низкую точность для сложных наборов данных (например, для больших наборов данных и наборов данных со сложными взаимодействиями переменных).

В настройке «случайного леса» многие деревья классификации и регрессии строятся с использованием случайно выбранных наборов, обучающих данных и случайных подмножеств переменных-предикторов для результатов моделирования. Результаты каждого дерева агрегированы, чтобы дать прогноз для каждого наблюдения [32, 33]. Следовательно, «случайный лес» часто обеспечивает

более высокую точность по сравнению с моделью с одним деревом решений, сохраняя при этом некоторые полезные качества моделей деревьев. «Случайные леса» неизменно предлагают одну из самых высоких точность прогнозов по сравнению с другими моделями при настройке классификации.

Основным преимуществом использования «случайного леса» для моделирования прогнозирования является возможность обрабатывать наборы данных с большим количеством переменных-предикторов, однако часто на практике количество предикторов, необходимых для получения прогнозов результатов, следует минимизировать для повышения эффективности. Например, вместо использования всех переменных, имеющихся в электронной медицинской карте, можно предпочесть использовать только подмножество наиболее важных переменных при разработке модели медицинского прогнозирования [34]. В прогнозном моделировании часто возникает интерес определить наиболее важные предикторы, которые следует включить в сокращенную и экономную модель. Это может быть достигнуто путем выбора переменных, при которых оптимальные предикторы определяются на основе статистических характеристик, таких как важность или точность. Разработка моделей прогнозирования с использованием выбора переменных может снизить нагрузку на сбор данных и повысить эффективность прогнозирования на практике [35]. Поскольку многие современные наборы данных имеют сотни или тысячи возможных предикторов, выбор переменных часто является необходимой частью разработки модели прогнозирования.

Выбор переменных в структуре «случайного леса» является важным аспектом для многих приложений в экспертных системах и приложениях. Таким образом, общая цель многих экспертных систем – помочь в принятии решений по сложной проблеме.

«Случайный лес» использует такие алгоритмы, как бэггинг и случайности признака.

Случайность признака. В обычном дереве решений, когда нужно разделить узел, мы рассматриваем каждый возможный признак и выбираем тот, который «сильнее» делит значения в узлах. В «случайном лесу» каждое дерево может делать выбор исключительно из случайного подмножества объектов. Из этого следует еще большая вариация между деревьями в модели. В конечном итоге более слабая корреляция между деревьями соответствует большему разнообразию.

Бэггинг. Деревья решений очень чувствительны к данным, на которых обучаются: небольшие изменения в наборе могут привести к значительно отличающимся древовидным структурам. «Случайный лес» использует это преимущество, позволяя каждому отдельному дереву произвольно выбирать данные с заменой, что приводит к различным деревьям.

Существует несколько методов выбора переменных при случайной классификации лесов. Многие пакеты **Randomforest** предоставляют процедуры случайного выбора переменных леса.

Инкрементальная кластеризация. Инкрементальная кластеризация – это задача разделения набора данных на k кластеров, в которых точки в одном кластере похожи, а точки в разных кластерах не похожи [36]. Контекст инкрементальной кластеризации выглядит следующим образом: для некоторых текущих кластеров инкрементальная кластеризация является однопроходной кластеризацией, цель которой – идентифицировать метку кластера для точек инкрементальных данных.

Инкрементальная кластеризация очень выгодна для динамических данных или потока данных (хранилища данных). Как правило, инкрементальная кластеризация комбинируется с несколькими процессами удаления и вставки. Учитывая набор кластеров, этап вставки направлен на идентификацию меток, новой точки данных на основе текущих кластеров. В некоторых случаях будут созданы новые кластеры или новые точки данных будут интегрированы с текущими кластерами.

В процессе удаления, если мы хотим удалить одну или несколько точек данных, нам необходимо реформировать кластеры, потому что эти операции могут повлиять на некоторые уже существующие кластеры. Для любого типа кластеризации в литературе предложены некоторые алгоритмы инкрементальной кластеризации, такие как Incremental k-means, Incremental DBSCAN (Density Based Spatial Clustering of Applications with Noise) или инкрементальная кластеризация графов [37]. Ключевая идея этих алгоритмов заключается в том, что нам необходимо идентифицировать ситуацию для каждого типа алгоритма для шага вставки и шага удаления.

Инкрементальная кластеризация решает проблему идентификации метки для нового объекта данных или обновления кластеров, когда мы удаляем точки в текущих кластерах. Эта проблема очень важна, когда мы занимаемся большими данными, набор данных которых слишком велик, чтобы поместиться в доступной памяти. Для каждого вида кластеризации в литературе предлагается несколько вариантов инкрементальной кластеризации.

Существует кластеризация на основе инкрементальной плотности (Incremental DBSCAN). Опираясь на концепцию кластеризации на основе плотности, Incremental DBSCAN может результативно прибавлять и удалять точки для текущих кластеров. Процесс добавления новой точки имеет несколько случаев (например, новая точка может быть шумом, новая точка будет добавлена в кластер, новая точка может объединить несколько кластеров). Для процесса удаления точка может быть точкой шума, может разбивать на некоторые кластеры или не влиять на текущие кластеры.

Также есть однопроходная инкрементальная кластеризация для большого набора данных на основе *k*-средних, которую обозначили как GenIC (Generalized Incremental Algorithm for Clustering). GenIC обновляет каждый центр с каждой новой точкой данных и объединяет кластеры только в конце генерации (то есть окна данных). С помощью обобщенного инкрементального алгоритма алгоритм GenIC может перемещать центр в списке центров, используя взвешенную сумму существующего центра и представленной новой точки.

Сущность алгоритма GenIC состоит в том, чтобы разбить поток данных на блоки или окна, как это заведено в алгоритмах потоковой передачи. Мы рассматриваем каждый блок из *n* точек данных как поколение и думаем, что «приспособленность» центра измеряется количеством назначенных ему точек. В целом наиболее приспособленные центры доживают до следующего поколения, но иногда выбираются новые центры, а старые центры уничтожаются. Алгоритм GenIC сравнивается с *k*-средними, показывает эффективность по времени выполнения и меньше зависит от выбора начальных центров, чем *k*-средние.

Предлагается вариант инкрементальной кластеризации *k*-средних. В алгоритме кластеры строятся постепенно, добавляя по одному центру кластера за один раз. Представлены новый двухфазный статический однопроходный алгоритм, а также динамический двухфазный однопроходный алгоритм, основанный на методе нечеткой кластеризации *C*-средних, которые демонстрируют высокую полезность [38]. Идея, лежащая в основе многоступенчатых методов, заключается в том, что оценка матрицы разделения и расположения центров кластеров может быть получена путем кластеризации выборки данных. Ожидается, что небольшая выборка дает быструю, но менее надежную оценку центров кластеров.

Это приводит к многоступенчатому подходу, который включает несколько этапов выборки (с заменой) данных и оценки матрицы членства для следующего этапа. Проведенные эксперименты показывают эффективность предложенного метода, схему инкрементальной кластеризации локальной плотности для поиска плотных подграфов в потоковых данных, то есть, когда данные поступают инкрементально (ILDC – incremental local density of clusters) [39].

Схема инкрементальной кластеризации захватывает избыточность в источнике потоковых данных, находя плотные подграфы, которые соответствуют заметным объектам и сценам. Процесс ILDC выполняет операции, такие как расширение кластера, добавление кластера и слияние кластера, на принципе подобия между определенными кластерами. ILDC показывает эффективность при использовании в приложениях для поиска изображений.

Далее рассмотрим алгоритм инкрементальной полууправляемой ансамблевой кластеризации, названный ISSCE (incremental semi-supervised clustering ensemble). Алгоритм ISSCE использует ограничения для обновления инкрементальных элементов. Разрабатывается процесс возрастающего выбора членов ансамбля на основе глобальной целевой функции и локальной целевой функции для удаления избыточных членов ансамбля.

Представлено улучшение ISSCE по сравнению с традиционными подходами к ансамблям полууправляемой кластеризации или с традиционными методами ансамбля кластеров на шести реальных наборах данных из репозитория машинного обучения UCI и на 12 реальных наборах данных профилей экспрессии [40]. В контексте классификации находят метку для нового объекта данных с помощью классификатора, обученного на предлагаемых данных. Проблема идентификации метки для нового объекта в инкрементальной кластеризации может рассматриваться аналогично контексту классификации.

Алгоритм IPLoM (Iterative Partitioning Log Mining). Алгоритм IPLoM разработан для кластеризации данных журнала. Он работает путем итеративного разбиения наборов журналов событий, используемых в качестве учебных образцов. На каждом этапе процесса разбиения результирующие разделы становятся ближе к содержанию журналов событий. В конце процесса разбиения алгоритм пытается обнаружить форматы строк, которые создали строки в каждом разделе. Эти обнаруженные разделы и форматы строк являются выходными данными алгоритма [41].

Алгоритм IPLoM проходит четыре этапа:

1) разделение по количеству токенов;
2) разделение по позиции токена;
3) разбиение поиском на биекцию;
4) извлечение шаблона журнала.

Этапы описаны более подробно ниже. Алгоритм предназначен для обнаружения всех возможных форматов строк в начальном наборе журналов событий. Алгоритм находит только форматы строк, поддержка которых превышает определенный порог, функция сокращения файла включена в алгоритм [42]. Функция сокращения файла работает, избавляясь от всех разделов, которые опускаются ниже порога поддержки файлов. Новое значение фиксируется в конце каждого шага разделения. Таким образом, мы способны производить только линейные форматы, которые соответствуют желаемому порогу поддержки файлов в конце алгоритма. Работа IPLoM без порога поддержки файлов является его состоянием по умолчанию.

Шаг 1. Разделение по величине событий. Журналы разбиты на различные кластеры по длине. В настоящих журналах может быть такое, что журналы,

принадлежащие одному шаблону, могут иметь переменную длину. В этом случае результат IPLoM следует обработать вручную.

Шаг 2. Разделение по позиции токена. На этом шаге каждый кластер содержит журналы схожей длины. Рассчитывая, что в кластере есть m журналов, длина которых равна n, этот кластер можно рассматривать как матрицу размером $m \times n$. Данный шаг сделан на предположении, что столбец с минимальным числом неповторимых слов (позиция разбитого слова) содержит константы. В итоге позиция разбитого слова применяется для разделения каждого кластера, то есть каждый сгенерированный кластер имеет одно и то же слово в позиции разбитого слова.

Шаг 3. Разбиение с помощью поиска по отображению. На данном шаге два столбца журналов выбираются для предстоящего разбиения на базе соотношения отображения между ними. Чтобы найти два столбца, определяется количество уникальных слов в каждом столбце и выбирается два столбца с чаще всего встречающимся числом слов. Существует четыре отношения отображения: 1-1, 1-M, M-1, M-M. В случае отношений 1-1 журналы содержат одинаковые отношения 1-1 в двух выбранных столбцах, разделенных на один и тот же кластер. Для отношений 1-M и M-1 мы должны сначала решить, содержит ли столбец стороны M константы или переменные. Если сторона M содержит константы, то столбец стороны M используется для разделения журналов в отношениях 1-M / M-1. В противном случае используется первая боковая колонка. Наконец, журналы в отношениях M-M делятся на один кластер.

Шаг 4. Извлечение шаблона журнала. Алгоритм IPLoM обрабатывает все кластеры, которые были созданы на прошлых шагах, и генерирует по одному шаблону журнала для каждого из них. Для каждого столбца в кластере подсчитывается число неповторимых слов. Если в столбце только одно неповторимое слово, оно считается неизменным. Иначе слова в столбце являются переменными и в выходных данных будут сменены знаком подстановки.

3. ПРОБЛЕМА АНАЛИЗА ЖУРНАЛОВ БЕЗОПАСНОСТИ

Файлы журнала безопасности содержат данные практически обо всех событиях, которые происходят в системе, независимо от уровня журнала. Для этого развернутая инфраструктура ведения журналов автоматически собирает, объединяет и хранит журналы, которые постоянно создаются большинством компонентов и устройств (например, веб-серверами, базами данных или межсетевыми экранами). Текстовые сообщения журнала обычно читабельны и привязываются к отметке времени, сообщая момент времени,

когда была создана запись журнала. Доступ к длительным данным журнала имеет огромное количество преимуществ, в особенности для больших компаний: журналы дают возможность проводить изучение прошедших событий, системные администраторы получают возможность отследить корни наблюдаемых проблем. Кроме того, журналы могут содействовать возврату системы к исправному состоянию, сбросить некорректные операции, вернуть данные, предупредить утрату информации и воспроизвести сценарии, которые приводят к неверным состояниям во время тестирования [43].

Главная трудность изучения журналов состоит в том, что инциденты обнаруживаются лишь постфактум. Также изучение журналов – это трудозатратная и ресурсоемкая задача, которая требует познания предметной области о системе. В настоящее время обнаружение дефектов в системах становится вероятным благодаря неизменному мониторингу системных журналов в реальном времени, т. е. online. Это позволяет вовремя проявлять реакцию на происшествия кибербезопасности [44] и снижает вызванные ими потери. Также индикаторы грядущего неверного поведения системы нередко можно отследить заблаговременно. Довольно раннее обнаружение таковых индикаторов и принятие соответствующих мер может предупредить определенные неисправности.

Но эта задача навряд ли вероятна для человека, так как данные журнала генерируются в больших объемах [45] и с большой скоростью. При рассмотрении больших корпоративных систем часто число каждый день создаваемых строк журнала исчисляется миллионами. К примеру, общедоступные журналы распределенной файловой системы Hadoop Distributed File System (HDFS) содержат свыше 4 млн строк журнала в день, а маленькие организации имеют дело с наивысшими показателями 22 тыс. событий за секунду. Разумеется, что это делает неосуществимым ручной анализ, и потому уместно применять методы машинного обучения [46], которые автоматом обрабатывают линии и распознают достойные внимания шаблоны, представляя их в сжатой форме.

Одним из способов изучения огромных объемов данных журнала является кластеризация. Данные журнала владеют определенными чертами, которые нужно учесть при разработке метода кластеризации. Во-первых, файл журнала обычно состоит из набора однострочных либо многострочных строк, перечисленных в определенном хронологическом порядке, который обычно подкрепляется отметкой времени сообщений журнала [47]. Сообщения могут быть очень структурированными (к примеру, перечень значений, разбитых запятыми), отчасти структурированными (к примеру, пары атрибут-значений), неструктурированными (к примеру, вольный текст случайной длины) либо смешанными, а также сообщения журнала время от времени включают иден-

тификаторы действий, относящихся к задаче, которая их сгенерировала. В данном случае достаточно просто извлечь трассировки журнала, другими словами, последовательности связанных строк журнала, и выполнить интеллектуальный анализ процессов. Остальные артефакты, время от времени включаемые в сообщения журнала, – это номера строк, индикатор уровня либо серьезности сообщения и статический идентификатор, указывающий на оператора, который создает сообщение.

Эти характеристики разрешают группировать системные журналы двумя различными методами. Во-первых, кластеризация некоторых строк журнала по схожести их сообщений дает обзор всех событий, которые происходят в системе. Во-вторых, кластеризация последовательностей сообщений журнала дает представление о базисной логике программы и открывает при другом варианте скрытые зависимости событий и элементов.

ЗАКЛЮЧЕНИЕ

В настоящей работе были рассмотрены наименее известные, но не менее эффективные и обладающие высокой точностью методы и алгоритмы кластеризации. Рассмотрены типы журналов событий. Особое внимание уделено журналам безопасности. Следует отметить, что обсуждаемые в статье методы и алгоритмы кластеризации могут быть применены для анализа журналов безопасности.

БЛАГОДАРНОСТИ

Авторы выражают глубокую благодарность д-ру техн. наук, профессору Белову Виктору Матвеевичу за ценные советы и замечания, высказанные при работе над статьей.

СПИСОК ЛИТЕРАТУРЫ

1. *Королев М.А.* Статистический словарь. – М.: Финансы и статистика, 1989. – 623 с.

2. *Воронцов К.В.* Алгоритмы кластеризации и многомерного шкалирования: курс лекций. – М.: МГУ, 2007.

3. *Jain A., Murty M.*, *Flynn P.* Data clustering: a review // ACM Computing Surveys. – 1999. – Vol. 31, iss 3. – P. 264–323.

4. *Котов А.*, *Красильников Н.* Кластеризация данных. – СПб.: СПбГУ ИТМО, 2006.

5. *Мандель И.Д.* Кластерный анализ. – М.: Финансы и статистика, 1988. – 176 с.

6. Прикладная статистика: классификация и снижение размерности / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин. – М.: Финансы и статистика, 1989. – 607 с.

7. MachineLearning.Ru. Информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных. – URL: www.machinelearning.ru (дата обращения: 04.03.2022).

8. *Чубукова И.А.* Курс лекций «DataMining» / Интернет-университет информационных технологий. – URL: www.intuit.ru/department/database/datamining (дата обращения: 04.03.2022).

9. *Farid D.M., Rahman M.Z., Rahman C.M.* Adaptive intrusion detection based on boosting and naïve Bayesian classifier // International Journal of Computer Applications. – 2011. – Vol. 24 (3). – P. 12–19.

10. *Лепский А.Е., Броневич А.Г.* Математические методы распознавания образов: курс лекций. – Таганрог, 2009. – URL: https://lepskiy.ucoz.ru/Posobie/MMPR_.pdf (дата обращения: 04.03.2022).

11. Интуит. Национальный открытый университет. Лекция 9: Методы классификации и прогнозирования. Деревья решений. – URL: http://www.intuit.ru/studies/courses/6/6/lecture/174 (дата обращения: 14.03.2022).

12. *Круглов В.В., Голунов Р.Ю.* Нечеткая логика и искусственные нейронные сети. – М.: Физматлит, 2001. – 224 с.

13. *Воронцов К.В.* Лекции по искусственным нейронным сетям. – 2007, 21 декабря. – URL: http://www.ccas.ru/voron/download/NeuralNets.pdf (дата обращения: 04.03.2022).

14. *Барский А.Б.* Нейронные сети: распознавание, управление, принятие решений. – М.: Финансы и статистика, 2004. – 176 с.

15. *Панченко Т.В.* Генетические алгоритмы / под ред. Ю.Ю. Тарасевича. – Астрахань: Астраханский университет, 2007. – 87 с.

16. CompoWiki. Журнал событий. – URL: https://wiki.compowiki.info/ЖурналСобытий (дата обращения: 04.03.2022).

17. Журналы событий Windows. – URL: https://eventlogxp.com/rus/essentials/windowseventlog.html (дата обращения: 04.03.2022).

18. Журнал регистрации событий информационной безопасности. – URL: https://safe-surf.ru/glossary/ru/849/ (дата обращения: 04.03.2022).

19. *Makanju A., Zincir-Heywood A.N., Milios E.E.* Clustering event logs using iterative partitioning // KDD '09: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – ACM, 2009. – P. 1255–1264. – DOI: 10.1145/1557019.1557154.

20. On vulnerability and security log analysis: a systematic literature review on recent trends / J. Svacina, J. Raffety, C. Woodahl, B. Stone, T. Cerny, M. Bures, D. Shin, K. Frajtak, P. Tisnovsky // RACS '20: Proceedings of the International Conference on Research in Adaptive and Convergent Systems. – ACM, 2020. – P. 175–180. – DOI: 10.1145/3400286.3418261.

21. Process mining and hierarchical clustering to help intrusion alert visualization / S.C. de Alvarenga, S. Barbon, R.S. Miani, M. Cukier, B.B. Zarpelão // Computers and Security. – 2018. – Vol. 73. – P. 474–491. – DOI: 10.1016/j.cose.2017.11.021.

22. *Alaba A., Maitanmi S., Ajayi O.* An ensemble of classification techniques for Intrusion detection systems // International Journal of Computer Science and Information Security. – 2019. – Vol. 17, N 11. – P. 24–33.

23. *Chauhan A., Mishra G., Kumar G.* Survey on data mining techniques in intrusion detection // International Journal of Scientific and Engineering Research. – 2011. – Vol. 2, iss. 7. – P. 1–4.

24. A multi-level intrusion detection method for abnormal network behavior / S.-Y. Ji, S. Choi, B.-K. Jeong, D.H. Jeong // Journal of Network and Computer Applications. – 2016. – Vol. 62. – P. 9–17.

25. *Onan A., Korukoğlu S., Bulut H.* A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification // Expert Systems with Applications. – 2016. – Vol. 62. – P. 1–16. – DOI: 10.1016/j.eswa.2016.06.005.

26. Implementation of naïve Bayes classification method for predicting purchase / F. Harahap, A.Y.N. Harahap, E. Ekadiansyah, R.N. Sari, R. Adawiyah, C.B. Harahap // 2018 6th International Conference on Cyber and IT Service Management (CITSM). – Parapat, Indonesia, 2018. – P. 1–5. – DOI: 10.1109/CITSM.2018.8674324.

27. Deep Learning techniques for traffic speed forecasting with side information / P. Farajiparvar, N. Hoseinzadeh, L.D. Han, A. Hedayatipour // 2020 IEEE Green Energy and Smart Systems Conference (IGESSC). – Long Beach, CA, 2020. – P. 1–5. – DOI: 10.1109/IGESSC50231.2020.9285132.

28. *Aklani S.A.* Metode fuzzy logic untuk evaluasi kinerja pelayanan perawat (Studi Kasus: RSIA Siti Hawa Padang) // Edik Informatika. – 2014. – Vol. 1, N 1. – P. 35–43.

29. Recognition of driving postures by contourlet transform and random forests / C.H. Zhao, B.L. Zhang, J. He, J. Lian // IET Intelligent Transport Systems. – 2012. – Vol. 6 (2). – P. 161–168.

30. *Probst P., Wright M.N., Boulesteix A.-L.* Hyperparameters and tuning strategies for random forest // WIREs Data Mining and Knowledge Discovery. – 2019. – Vol. 9. – P. e1301.

31. Applying a random forest method approach to model travel mode choice behavior / L. Cheng, X. Chen, J. De Vos, X. Lai, F. Witlox // Travel Behaviour and Society. – 2019. – Vol. 14. – P. 1–10.

32. GIS-based landslide susceptibility evaluation using a novel hybrid integration approach of bivariate statistical based random forest method / W. Chen, X. Xie, J. Peng, H. Shahabi, H. Hong, D.T. Bui, Z. Duan, S. Li, A-X. Zhu // Catena. – 2018. – Vol. 164. – P. 135–149.

33. Identifying core driving factors of urban land use change from global land cover products and POI data using the random forest method / H. Wu, A. Lin, X. Xing, D. Song, Y. Li // International Journal of Applied Earth Observation and Geoinformation. – 2021. – Vol. 103. – P. 102475.

34. *Cai Y., Lin H., Zhang M.* Mapping paddy rice by the object-based random forest method using time series Sentinel-1/Sentinel-2 data // Advances in Space Research. – 2019. – Vol. 64 (11). – P. 2233–2244.

35. Prediction of consumer behaviour using random forest algorithm / H. Valecha, A. Varma, I. Khare, A. Sachdeva, M. Goyal // 2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON). – Gorakhpur, India, 2018. – P. 1–6. – DOI: 10.1109/UPCON.2018.8597070.

36. *Кутуков Д.С.* Применение методов кластеризации для обработки новостного потока // Технические науки: проблемы и перспективы: материалы I Международной научной конференции. – СПб.: Реноме, 2011. – С. 77–83. – URL: https://moluch.ru/conf/tech/archive/2/207/ (дата обращения: 09.03.2022).

37. *Kailing K., Kriegel H.-P., Kröger P.* Density-connected subspace clustering for high-dimensional data // Proceedings of the 4th SIAM International Conference on Data Mining (SDM). – Philadelphia, PA, 2004. – P. 246–257.

38. *Braun R.K.*, *Kaneshiro R.* Exploiting topic pragmatics for new event detection in TDT-2004 // DARPA Topic Detection and Tracking Workshop. – Gaithersburg, 2004.

39. *Peters M.*, *Zaki M.J.* Click: clustering categorical data using K-partite maximal cliques / Computer Science Department Rensselaer Polytechnic Institute. – Troy, NY, 2004. – 31 p.

40. Clustering uncertain data based on probability distribution similarity / B. Jiang, J. Pei, Y. Tao., X. Lin // IEEE Transactions on Knowledge and Data Engineering. – 2013. – Vol. 25 (4). – P. 751–763. – DOI: 10.1109/TKDE.2011.221.

41. *Makanju A., Zincir-Heywood A.N., Milios E.E.* A lightweight algorithm for message type extraction in system application logs // IEEE Transactions on Knowledge and Data Engineering. – 2012. – Vol. 24 (11). – P. 1921–1936. – DOI: 10.1109/TKDE.2011.138.

42. *Makanju A., Zincir-Heywood A.N., Milios E.E.* Clustering event logs using iterative partitioning // ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 09). – ACM, 2009. – P. 1255–1263.

43. *Oliner A., Ganapathi A., Xu W.* Advances and challenges in log analysis: logs contain a wealth of information for help in managing systems // ACM Queue. – 2011. – Vol. 9 (12). – DOI: 10.1145/2076796.2082137.

44. Best practices for incident response. – 2020, September 3. – URL: https://www.securitymagazine.com/articles/93235-best-practices-for-incident-response (accessed: 09.03.2022).

45. Operational-log analysis for big data systems: challenges and solutions / A. Miranskyy, A. Hamou-Lhadj, E. Cialini, A. Larsson // IEEE Software. – 2016. – Vol. 33 (2). – P. 52–59. – DOI: 10.1109/MS.2016.33.

46. HDFS Architecture / The Apache Software Foundation. – URL: https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html (accessed: 09.03.2022).

47. *Brownlee J.* A tour of machine learning algorithms. – 2019, August 12. – URL: https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/ (accessed: 09.03.2022).

Сидорова Диана Николаевна, ассистент кафедры защиты информации Новосибирского государственного технического университета. В настоящее время специализируется в области информационной безопасности. E-mail: d.sidorova.2013@corp.nstu.ru

Пивкин Евгений Николаевич, кандидат технических наук, руководитель направления отдела защиты информации департамента безопасности ПАО АКБ «Связь-Банк». Основное направление научных исследований – применение математических методов в различных областях науки, техники, общества. Автор более 80 публикаций. E-mail:evpiv@yandex.ru

DOI: 10.17212/2782-2230-2022-1-41-60

Algorithms and methods of data clustering in the analysis of information security event logs*

D.N. Sidorova[1], E.N. Pivkin[2]

[1] *Novosibirsk State Technical University, 20 K. Marx Prospekt, Novosibirsk, 630073, Russian Federation, assistant of the Department of Information Security. E-mail: d.sidorova.2013@corp.nstu.ru*
[2] *PJSC JSCB "Svyaz-Bank", 31/7 Novoryazanskaya Street, Moscow, 105066, Russian Federation, candidate of technical sciences, E-mail: evpiv@yandex.ru*

Security event log files give an idea of the state of the information system and allow you to find anomalies in user behavior and cybersecurity incidents. The existing event logs (application, system, security event logs) and their division into certain types are considered. But automated analysis of security event log data is difficult because it contains a large amount of unstructured data that has been collected from various sources. Therefore, this article presents and describes the problem of analyzing information security event logs.
And to solve this problem, new and not particularly studied methods and algorithms for data clustering were considered, such as Random forest (random forest), incremental clustering, IPLoM algorithm (Iterative Partitioning Log Mining - iterative analysis of the partitioning log). The Random forest algorithm creates decision trees for data samples, after which it is provided with a forecast for each sample, and the best solution is selected by voting. This method reduces overfitting by averaging the scores. The algorithm is also used in such types of problems as regression and classification. Incremental clustering defines clusters as groups of objects that belong to the same class or concept, which is a specific set of pairs. When clusters are defined, they can overlap, allowing for a degree of "fuzziness for samples" that lie at the boundaries of different clusters. The IPLoM algorithm uses the unique characteristics of log messages to iteratively partition the log, which helps to extract message types efficiently.

Keywords: algorithms, methods, data clustering, information security, random forest, incremental clustering, iterative partitioning log analysis, event logs

REFERENCES

1. Korolev M.A. *Statisticheskii slovar'* [Statistical dictionary]. Moscow, Finansy i statistika Publ., 1989. 623 p.
2. Vorontsov K.V. *Algoritmy klasterizatsii i mnogomernogo shkalirovaniya*: kurs lektsii [Clustering and multidimensional scaling algorithms. Lecture course]. Moscow State University, 2007.
3. Jain A., Murty M., Flynn P. Data clustering: a review. *ACM Computing Surveys*, 1999, vol. 31, iss. 3, pp. 264–323.

* *Received 07 February 2022.*

4. Kotov A., Krasil'nikov N. *Klasterizatsiya dannykh* [Data clustering]. St. Petersburg, ITMO University, 2006.

5. Mandel' I.D. *Klasternyi analiz* [Cluster analysis]. Moscow, Finansy i statistika Publ., 1988. 176 p.

6. Aivazyan S.A., Bukhshtaber V.M., Enyukov I.S., Meshalkin L.D. *Prikladnaya statistika: klassifikatsiya i snizhenie razmernosti* [Applied statistics: classification and dimensionality reduction]. Moscow, Finansy i statistika Publ., 1989. 607 p.

7. *MachineLearning.Ru*. Information and analytical resource dedicated to machine learning, pattern recognition and data mining. (In Russian). Available at: www.machinelearning.ru (accessed 04.03.2022).

8. Chubukova I.A. *Kurs lektsii "DataMining"* [Lecture course "Data Mining"]. Internet University of Information Technologies. Available at: www.intuit.ru/department/database/datamining (accessed 04.03.2022).

9. Farid D.M., Rahman M.Z., Rahman C.M. Adaptive intrusion detection based on boosting and naïve Bayesian classifier. *International Journal of Computer Applications*, 2011, vol. 24 (3), pp. 12–19.

10. Lepskiy A.E., Bronevich A.G. *Matematicheskie metody raspoznavaniya obrazov*: kurs lektsii [Mathematical methods for pattern recognition]. Taganrog, 2009. Available at: https://lepskiy.ucoz.ru/Posobie/MMPR_.pdf (accessed 04.03.2022).

11. Intuit. National Open University. Lektsiya 9: *Metody klassifikatsii i prognozirovaniya. Derev'ya reshenii* [Lecture 9: Classification and forecasting methods. decision trees]. Available at: http://www.intuit.ru/studies/courses/6/6/lecture/174 (accessed 14.03.2022).

12. Kruglov V.V., Golunov R.Yu. *Nechetkaya logika i iskusstvennye neironnye seti* [Fuzzy logic and artificial neural networks]. Moscow, Fizmatlit Publ., 2001. 224 p.

13. Vorontsov K.V. *Lektsii po iskusstvennym neironnym setyam* [Lectures on artificial neural networks], 2007, December 21. Available at: http://www.ccas.ru/voron/download/NeuralNets.pdf (accessed 04.03.2022).

14. Barskii A.B. *Neironnye seti: raspoznavanie, upravlenie, prinyatie reshenii* [Neural networks: recognition, control, decision making]. Moscow, Finansy i statistika Publ., 2004. 176 p.

15. Panchenko T.V. *Geneticheskie algoritmy* [Genetic algorithms]. Astrakhan, Astrakhanskii universitet Publ., 2007. 87 p.

16. CompoWiki. *Zhurnal sobytii* [CompoWiki. Event log]. Available at: https://wiki.compowiki.info/EventLog (accessed 04.03.2022).

17. *Zhurnaly sobytii Windows* [Windows event log]. Available at: https://eventlogxp.com/rus/essentials/windowseventlog.html (accessed 04.03.2022).

18. *Zhurnal registratsii sobytii informatsionnoi bezopasnosti* [Information security event log]. Available at: https://safe-surf.ru/glossary/ru/849/ (accessed 04.03.2022).

19. Makanju A., Zincir-Heywood A.N., Milios E.E. Clustering event logs using iterative partitioning. *KDD '09: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2009, pp. 1255–1264. DOI: 10.1145/1557019.1557154.

20. Svacina J., Raffety J., Woodahl C., Stone B., Cerny T., Bures M., Shin D., Frajtak K., Tisnovsky P. On vulnerability and security log analysis: a systematic literature review on recent trends. *RACS '20: Proceedings of the International Conference on Research in Adaptive and Convergent Systems*. ACM, 2020, pp. 175–180. DOI: 10.1145/3400286.3418261.

21. Alvarenga S.C. de, Barbon S., Zarpelão B.B., Miani R.S., Cukier M. Process mining and hierarchical clustering to help intrusion alert visualization. *Computers and Security*, 2018, vol. 73, pp. 474–491. DOI: 10.1016/j.cose.2017.11.021.

22. Alaba A., Maitanmi S., Ajayi O. An ensemble of classification techniques for Intrusion detection systems. *International Journal of Computer Science and Information Security*, 2019, vol. 17, no. 11, pp. 24–33.

23. Chauhan A., Mishra G., Kumar G. Survey on data mining techniques in intrusion detection. *International Journal of Scientific and Engineering Research*, 2011, vol. 2, iss. 7, pp. 1–4.

24. Ji S.-Y., Choi S., Jeong B.-K., Jeong D.H. A multi-level intrusion detection method for abnormal network behavior. *Journal of Network and Computer Applications*, 2016, vol. 62, pp. 9–17.

25. Onan A., Korukoğlu S., Bulut H. A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. *Expert Systems with Applications*, 2016, vol. 62, pp. 1–16. DOI: 10.1016/j.eswa.2016.06.005.

26. Harahap F., Harahap A.Y.N., Ekadiansyah E., Sari R.N., Adawiyah R., Harahap C.B. Implementation of naïve Bayes classification method for predicting purchase. *2018 6th International Conference on Cyber and IT Service Management (CITSM)*, Parapat, Indonesia, 2018, pp. 1–5. DOI: 10.1109/CITSM.2018.8674324.

27. Farajiparvar P., Hoseinzadeh N., Han L.D., Hedayatipour A. Deep Learning techniques for traffic speed forecasting with side information. *2020 IEEE Green Energy and Smart Systems Conference (IGESSC)*, Long Beach, CA, 2020, pp. 1–5. DOI: 10.1109/IGESSC50231.2020.9285132.

28. Aklani S.A. Metode fuzzy logic untuk evaluasi kinerja pelayanan perawat (Studi Kasus: RSIA Siti Hawa Padang). *Edik Informatika*, 2014, vol. 1, no. 1, pp. 35–43.

29. Zhao C.H., Zhang B.L., He J., Lian J. Recognition of driving postures by contourlet transform and random forests. *IET Intelligent Transport Systems*, 2012, vol. 6 (2), pp. 161–168.

30. Probst P., Wright M.N., Boulesteix A.-L. Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery*, 2019, vol. 9, p. e1301.

31. Cheng L., Chen X., Cheng L., De Vos J., Witlox F., Lai X., Witlox F., Witlox F. Applying a random forest method approach to model travel mode choice behavior. *Travel Behaviour and Society*, 2019, vol. 14, pp. 1-10.

32. Chen W., Xie X., Peng J., Shahabi H., Hong H., Bui D.T., Duan Z., Li S., Zhu A-X. GIS-based landslide susceptibility evaluation using a novel hybrid integration approach of bivariate statistical based random forest method. *Catena*, 2018, vol. 164, pp. 135–149.

33. Wu H., Lin A., Xing X., Song D., Li Y. Identifying core driving factors of urban land use change from global land cover products and POI data using the random forest method. *International Journal of Applied Earth Observation and Geoinformation*, 2021, vol. 103, p. 102475.

34. Cai Y., Lin H., Zhang M. Mapping paddy rice by the object-based random forest method using time series Sentinel-1/Sentinel-2 data. *Advances in Space Research*, 2019, vol. 64 (11), pp. 2233–2244.

35. Valecha H., Varma A., Khare I., Sachdeva A., Goyal M. Prediction of consumer behaviour using random forest algorithm. *2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, Gorakhpur, India, 2018, pp. 1–6. DOI: 10.1109/UPCON.2018.8597070.

36. Kutukov D.S. [Application of clustering methods for news flow processing]. *Tekhnicheskie nauki: problemy i perspektivy*: materialy I Mezhdunarodnoi nauchnoi konferentsii [Technical sciences: problems and prospects: materials of the I International scientific conference], St. Petersburg, Renome Publ., 2011, pp. 77–83. (In Russian). Available at: https://moluch.ru/conf/tech/archive/2/207/ (accessed 09.03.2022).

37. Kailing K., Kriegel H.-P., Kröger P. Density-connected subspace clustering for high-dimensional data. *Proceedings of the 4th SIAM International Conference on Data Mining (SDM)*, Philadelphia, PA, 2004, pp. 246–257.

38. Braun R.K., Kaneshiro R. Exploiting topic pragmatics for new event detection in TDT-2004. *DARPA Topic Detection and Tracking Workshop*, Gaithersburg, 2004.

39. Peters M., Zaki M.J. *Click: clustering categorical data using K-partite maximal cliques*. Computer Science Department Rensselaer Polytechnic Institute, Troy, NY, 2004. 31 p.

40. Jiang B., Pei J., Tao Y, Lin X. Clustering uncertain data based on probability distribution similarity. *IEEE Transactions on Knowledge and Data Engineering*, 2013, vol. 25 (4), pp. 751–763. DOI: 10.1109/TKDE.2011.221.

41. Makanju A., Zincir-Heywood A.N., Milios E.E. A lightweight algorithm for message type extraction in system application logs. *IEEE Transactions on Knowledge and Data Engineering*, 2012, vol. 24 (11), pp. 1921–1936. DOI: 10.1109/TKDE.2011.138.

42. Makanju A., Zincir-Heywood A.N., Milios E.E. Clustering event logs using iterative partitioning. *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 09)*, ACM, 2009, pp. 1255–1263.

43. Oliner A., Ganapathi A., Xu W. Advances and challenges in log analysis: logs contain a wealth of information for help in managing systems. *ACM Queue*, 2011, vol. 9 (12). DOI: 10.1145/2076796.2082137.

44. *Best practices for incident response*. 2020, September 3. Available at: https://www.securitymagazine.com/articles/93235-best-practices-for-incident-response (accessed 09.03.2022).

45. Miranskyy A., Hamou-Lhadj A., Cialini E., Larsson A. Operational-log analysis for big data systems: challenges and solutions. *IEEE Software*, 2016, vol. 33 (2), pp. 52–59. DOI: 10.1109/MS.2016.33.

46. *HDFS Architecture*. The Apache Software Foundation. Available at: https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html (accessed 09.03.2022).

47. Brownlee J. *A tour of machine learning algorithms*. 2019, August 12. Available at: https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/ (accessed 09.03.2022).

Для цитирования:

Сидорова Д.Н., Пивкин Е.Н. Алгоритмы и методы кластеризации данных в анализе журналов событий информационной безопасности // Безопасность цифровых технологий. – 2022. – № 1 (104). – С. 41–60. – DOI: 10.17212/2782-2230-2022-1-41-60.

For citation:

Sidorova D.N., Pivkin E.N. Algoritmy i metody klasterizatsii dannykh v analize zhurnalov sobytii informatsionnoi bezopasnosti [Algorithms and methods of data clustering in the analysis of information security event logs]. *Bezopasnost' tsifrovykh tekhnologii = Digital Technology Security*, 2022, no. 1 (104), pp. 41–60. DOI: 10.17212/2782-2230-2022-1-41-60.

БЕЗОПАСНОСТЬ ЦИФРОВЫХ ТЕХНОЛОГИЙ. – 2022. – № 1 (104). – 61–84

МЕТОДЫ И СИСТЕМЫ ЗАЩИТЫ ИНФОРМАЦИИ, ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ

УДК 004.056 DOI: 10.17212/2782-2230-2022-1-61-84

ВОЗМОЖНОСТИ АНАЛИЗА НОМИНАТИВНЫХ ПРИЗНАКОВ В ЗАДАЧАХ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ*

В.Е. ХИЦЕНКО[1], Н.А. ФЕДОТОВ[2]

[1] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, кандидат технических наук, доцент. E-mail: xicenko@corp.nstu.ru*

[2] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, магистрант кафедры вычислительной техники. E-mail: nicoss.fd@yandex.ru*

В статье на различных примерах демонстрируются и обсуждаются возможности проверки гипотез и применения информационных мер для выявления и оценки силы связи номинативных признаков в задачах классификации при анализе информационной безопасности. Основной вид представления исходных данных в этой шкале – это таблица сопряженности номинативных признаков или таблица «объект – признак», из которой могут быть получены частоты совпадения категорий признаков и, собственно, таблица сопряженности. По этой таблице несложно проверить гипотезу о независимости или однородности признаков. Рассмотрен альтернативный подход к этому анализу на основе статистики Кульбака, представляющей собой среднюю различающую информацию в пользу гипотезы о зависимости признаков. В частных случаях практический интерес представляет гипотеза о симметрии квадратных таблиц, которая также может быть проверена на основе информационных мер и критериев. Показан пример обработки дихотомических данных типа «да – нет» по критерию Кокрена. В работе обсуждаются пути измерения силы связи признаков и различные информационные характеристики в виде относительного уменьшения энтропии одного признака при известном другом или в виде средневзвешенного количества информации, приходящегося на различные категории признака. Эти меры полезны для сравнительного анализа признаков в задачах принятия решений. Используются показатель информативности Шеннона, дивергенция Кульбака – Лейблера, Дженсена – Шеннона и мера попарного различения классов эффективности защиты по законам распределения соответствующих им категорий признака. Последовательно сопоставляются классические процедуры проверки гипотез и подходы на основе информационных характеристик. Рассмотренные в работе методы и примеры охватывают многие актуальные задачи информационной безопасности, ассоциированные с номинативными признаками.

* Статья получена 10 февраля 2022 г.

Ключевые слова: статистические методы, защита информации, проверка гипотез о сопряженности качественных признаков, критерии Кокрена, критерии Кульбака, проверка симметрии таблиц, меры связи признаков, информационный подход к анализу связи признаков, показатель информативности Шеннона, дивергенция Кульбака – Лейблера и Дженсена – Шеннона

ВВЕДЕНИЕ

Номинативные признаки – это по сути наименования объектов. Например, способы защиты, уровни доступа, должности, номера отдела, типы программного обеспечения, виды вторжений, протоколы, техники обнаружения атак и т. п. Различие между признаками в такой слабой шкале не может быть измерено количественно. Однако связи между двумя и более такими признаками могут быть выявлены с помощью анализа таблицы сопряженности признаков, в ячейках которой фиксируются частоты или факты совпадения категорий (градаций) признаков в процессе наблюдения. В статье сравниваются два подхода к анализу таблиц: классический и информационный. Особый интерес представляют различные меры для оценки силы связи признаков, оценки их информативности и возможности этих мер для классификации и принятия решений в сфере информационной безопасности.

1. КЛАССИЧЕСКИЙ АНАЛИЗ СОПРЯЖЕННОСТИ ПРИЗНАКОВ

Обсудим несколько подходов к измерению сопряженности признаков. Исходные данные представляются в виде таблицы частот совпадения категорий исследуемых признаков.

Таблица 1

Table 1

Таблица сопряженности двух признаков

Contingency table of two features

Признак 1	Признак 2				
	1	*2*	…	*l*	Сумма
1	n_{11}	n_{12}	…	n_{1l}	$n_{1\cdot}$
2	n_{21}	n_{22}	…	n_{2l}	$n_{2\cdot}$
⋮	⋮	⋮	⋮	⋮	⋮
k	n_{k1}	n_{k2}	…	n_{kl}	$n_{k\cdot}$
Сумма	$n_{\cdot 1}$	$n_{\cdot 2}$	…	$n_{\cdot l}$	$n_{\cdot\cdot}$

Количество категорий первого признака (число строк) равно k. Второй признак представлен l категориями-столбцами; n_{ij} – частота совпадений i-й строки и j-го столбца; $n_{i\cdot}$ и $n_{\cdot j}$ – суммы частот i-й строки и j-го столбца соответственно; $n_{\cdot\cdot} = n$ – объем выборки.

Разные группы людей или объектов могут отличаться и классифицироваться именно по распределению частот. В другом случае мы имеем одну группу объектов и располагаем частотами проявления одного признака в разных условиях или, что характерно для этой шкалы, частотами проявления разных признаков при фиксированных условиях. Однако принадлежность к группе можно рассматривать как еще один номинативный признак, как столбец «переменная группирования» таблицы «объект-признак», строки которой соответствуют разным объектам.

Метод давно используют для выявления связи двух номинативных признаков, представленных таблицами сопряженности [1, 2]. Для этого находят ожидаемые частоты, которые соответствуют полному отсутствию связи между строками и столбцами (гипотеза H_0) и вычисляют некоторую меру отклонения частот от ожидаемых.

Рассмотрим метод на иллюстративном примере. В табл. 2 показаны частоты сопряженности признака 1 (разные группы специалистов) и признака 2 (мнения об уровне защиты). Ранговую шкалу с небольшим количеством категорий (скажем, низкий, средний, высокий) обычно также относят к номинативной.

Таблица 2

Table 2

Анализ различия мнений сотрудников об уровне защиты

The analysis of distinction of opinions of employees about protection level

Признак 1	Признак 2			
	Низкий	Средний	Высокий	$n_{i\cdot}$
Группа 1	24	7	7	38
Группа 2	76	38	70	184
Группа 3	69	32	82	183
Группа 4	27	9	55	91
$n_{\cdot j}$	196	86	214	$n_{\cdot\cdot} = 496$

Найдем ожидаемые частоты, которые получились бы при полном отсутствии связи этих двух признаков, то есть зависимости мнения от принадлежности к группе (гипотеза H_0). Эти частоты находятся по формуле

$$n_{ij}^0 = \frac{n_{i.} n_{.j}}{n_{..}}. \quad (1)$$

При справедливости H_0 статистика (2) распределена по закону χ-квадрат с параметром $(k-1)(l-1)$:

$$\chi^2_{\text{эмп}} = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(n_{ij}^0 - n_{ij}\right)^2}{n_{ij}^0}. \quad (2)$$

В данном примере имеем $\chi^2_{\text{эмп}} = 24{,}93$. Достигнутая значимость составляет менее 0,0004. Это вероятность получить такое или даже большее значение $\chi^2_{\text{эмп}}$ при справедливости H_0. Эту гипотезу отклоняем. Признаки явно сопряжены. Группы радикально отличаются во мнениях. Мы не можем считать это случайным, и выводы из анализа таблицы будут объективны и полезны. Примечательно, что гипотеза H_0 предполагает еще и однородность признака 1 относительно признака 2. Это означает, что все частоты принадлежности к группам при различных мнениях (признак 2) относятся к одной генеральной совокупности.

В сущности, та же статистика (2) традиционно используется для сравнения законов распределения двух признаков. Эта ситуация возникает, когда таблица сопряженности имеет две строки (например, филиал 1 и филиал 2). Столбцы определяют частоты предпочтения шести способов защиты (табл. 3).

Таблица 3

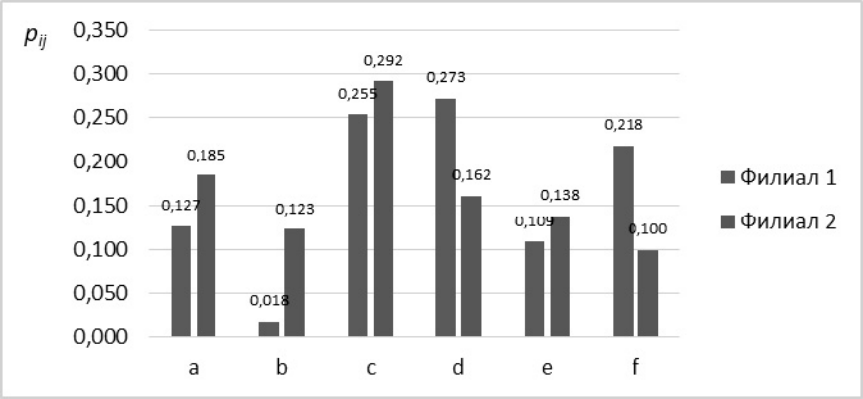
Table 3

Частоты предпочтения видов защиты в разных филиалах

Frequencies of preference for types of protection in different branches

Филиалы	Виды защиты						$n_{i.}$
	a	*b*	*c*	*d*	*e*	*f*	
Филиал 1	7	1	14	15	6	12	55
Филиал 2	24	16	38	21	18	13	130
$n_{.j}$	31	17	52	36	24	25	$n_{..} = 185$

Переходя к относительным частотам предпочтений $p_{ij} = n_{ij} / n_{i}.$, построим гистограммы (рисунок).



Гистограммы предпочтений способов защиты

Histograms of protection methods preferences

Визуальное сопоставление гистограмм не дает уверенности в существенном их различии. Однако проверка гипотезы H_0 о равенстве законов распределения предпочтений дает $\chi^2_{\text{эмп}} = 12,29$. Достигнутая значимость составляет 0,031. Гипотезу H_0 следует отклонить с риском ошибки около 3 % и исследовать причины различия предпочтений. Этот анализ послужит математической основой организационных или технических решений и выводов.

2. ИНФОРМАЦИОННЫЙ КРИТЕРИЙ СОПРЯЖЕННОСТИ

При этом подходе к анализу используют те же таблицы сопряженности признаков (табл. 1), однако затем применяют информационные меры [3], которые являются более гибким инструментом анализа по сравнению со статистикой (2).

Обозначим $p_{ij} = n_{ij} / n..$, $p_{i}. = n_{i}. / n..$, $p.{}_{j} = n.{}_{j} / n..$ эмпирические вероятности попадания в клетки таблицы. Для проверки гипотезы о независимости признаков H_0: «$p_{ij} = p_{i.}p_{.j}$ для всех клеток» против двусторонней альтернативы H_1: «$p_{ij} \neq p_{i.}p_{.j}$ хотя бы для одной клетки» вычисляют статистику,

представляющую собой удвоенную среднюю различающую информацию в пользу H_1 против H_0:

$$2I(H_1 : H_0) = 2\sum_{i=1}^{k}\sum_{j=1}^{l} p_{ij} \ln \frac{p_{ij}}{p_{i.}p_{.j}} = 2\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} \ln \frac{n_{..}n_{ij}}{n_{i.}n_{.j}} . \quad (3)$$

Статистика $I(H_1 : H_0)$ может быть представлена в виде (4), соответствующем разности логарифмов наблюдаемых и ожидаемых частот при справедливости H_0:

$$I(H_1 : H_0) = \sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} \ln n_{ij} - \sum_{i=1}^{k} n_{i.} \ln n_{i.} - \sum_{j=1}^{l} n_{.j} \ln n_{.j} + n \ln n . \quad (4)$$

Доказано, что при выполнении H_0 статистика (3) имеет в асимптотике то же χ^2-распределение с $(k - 1)(l - 1)$ степенями свободы. Если встречаются несколько равных нулю частот, то соответствующее слагаемое в (3) берем равным нулю и из полученной статистики вычитаем число таких слагаемых [1, с. 445].

Был реализован эксперимент по сравнению предпочтений альтернативных служебных клавиш. Исследование проводилось в рамках формирования клавиатурных портретов пользователей для задач идентификации и аутентификации [4]. В табл. 4 показаны сопряженности частот для пяти пользователей.

Таблица 4

Table 4

Частоты использования служебных клавиш для пяти пользователей

Service keys usage frequencies for five users

№ п/п	Левый Shift	Правый Shift	CapsLock	Левый Ctrl	Правый Ctrl	Backspace	Delete	$n_{i.}$
1	485	273	8	2	0	520	0	1288
2	1035	16	0	14	6	1639	0	2710
3	383	225	6	13	17	373	39	1056
4	338	372	0	5	11	27	304	1057
5	470	1	518	7	6	8	84	1094
$n_{.j}$	2711	887	532	41	40	2567	427	7205

Статистика (3) с учетом поправки оказалась огромной и составила 5757,4. Напомним, что это средняя информация в пользу H_1. Вероятность получить такое значение при справедливости H_0 равна нулю. Эти пять пользователей радикально различны по манере работы на клавиатуре.

Понятно, что при обычной проверке сопряженности признаков по критерию (2) результат получился бы таким же бесспорным: очевидно различие почерка.

Даже если повторить проверку для визуально более близких по почерку пользователей 1, 3 и 4, результат не изменится. Статистика (3) уменьшилась до $2I$ = 1152,7, но достигнутая значимость практически нулевая. Статистика (2), естественно, тоже очень велика, $\chi^2_{\text{эмп}}$ = 970,3.

Преимущество информационного подхода в возможности анализировать не только таблицы с небольшим числом нулевых ячеек, но и таблицы с тремя и более входами, то есть позволяющие исследовать условные независимости и взаимодействие более двух признаков. Это непараметрический, не предполагающий нормальность признаков дисперсионный анализ на основе частот сопряженности.

3. КРИТЕРИЙ КОКРЕНА (COCHRAN)

Этот метод применяется для анализа связи признаков, если мера сопряженности имеет двузначный характер типа «да – нет», «за – против» и измеряется на одной группе объектов в разных экспериментальных условиях или на разных, но одинаковых по численности группах [5]. Такая наиболее слабая среди номинативных шкала называется дихотомической. По сути, это выявление неоднородности наборов двоичных данных.

Рассмотрим метод на примере. Пять сотрудников проверяли на компетентность с помощью шести тестовых заданий. Результаты проверки в табл. 5. Выполнение отмечалось знаком «1», невыполнение – «0». Имеем гипотезу H_0: «все тесты одинаково трудны» против H_1: «есть различия».

Суммы элементов строк и столбцов обозначены здесь ΣX_R и ΣX_C соответственно. Находим Q-статистику [5, 6] по формуле

$$Q = \frac{(k-1)\left[k\sum\left(\sum X_C\right)^2 - \left(\sum\left(\sum X_C\right)\right)^2\right]}{k\sum\left(\sum X_C\right) - \sum\left(\sum X_C\right)^2} = \frac{(6-1)[6\cdot 57 - 17^2]}{6\cdot 17 - 59} = 6{,}163, \quad (5)$$

где k – число тестов (столбцов).

Таблица 5

Table 5

Иллюстрация метода Кокрена

Cochran's method illustration

Сотрудники	Тест 1	Тест 2	Тест 3	Тест 4	Тест 5	Тест 6	ΣX_R	$(\Sigma X_R)^2$
1	1	1	0	0	1	1	4	16
2	1	0	0	1	0	1	3	9
3	0	0	1	0	1	1	3	9
4	1	1	1	0	1	0	4	16
5	1	0	0	0	1	1	3	9
ΣX_C	4	2	2	1	4	4	17	59
$(\Sigma X_C)^2$	16	4	4	1	16	16	57	

Статистика Q при справедливости H_0 асимптотически распределена по закону χ^2 с параметром $k-1$. Достигнутая значимость 0,29 не так уж мала. Нет оснований считать, что тесты различаются и результаты тестирования необъективны. Если в результате проверки мы вынуждены отклонить H_0, встает задача совершенствования тестов.

После этого мы можем аналогично проверить гипотезу о близости компетентности сотрудников, просто поменяв местами строки и столбцы в табл. 3. В случае отклонения этой гипотезы мы обоснованно выявим наиболее и наименее компетентных сотрудников.

В результате такой проверки по транспонированной табл. 5 было получено $Q = 3{,}20$ со значимостью 0,525. Различий компетентности не обнаружено.

К сожалению, при ограниченных размерах таблиц χ^2-аппроксимация распределения статистики Q ненадежна. В таких ситуациях применяют статистическое моделирование огромного числа случайно сформированных таблиц. Затем находят долю таблиц, у которых Q достигает полученную в эксперименте. Это и есть оценка Монте-Карло достигнутой значимости, равная 0,273 для этого примера. Современные программы статистической обработки данных [6] предусматривают такое моделирование.

4. ПРОВЕРКА СИММЕТРИИ ТАБЛИЦ

Проверка симметрии таблиц возникает и может быть реализована в квадратных таблицах $k \times k$ произвольного размера. Ясно, что k – это число категорий одного и того же признака, но зафиксированного в разных условиях (например, до и после какого-то события, стажировки, изменения правил, законодательства, уровня защиты).

Гипотеза симметрии H_0 есть предположение о том, что клетки, симметричные главной диагонали, содержат равные частоты. Это значит, что число людей, изменивших мнение с категории i на категорию j, равно числу людей, поступивших прямо противоположно. Следовательно, H_0 : «$p_{ij} = p_{ji}$» против H_1 : «$p_{ij} \neq p_{ji}$ хотя бы для одной пары (i, j)».

Для проверки мы должны просуммировать квадраты разностей пар частот, симметричных относительно этой диагонали, деленные на сумму этих же частот. Желательно, чтобы число клеток с частотами менее трех не превышало пятой части всех клеток таблицы [1, 6].

Статистика $\chi^2_{\text{сим}}$ при справедливости гипотезы H_0 о симметрии распределена по закону χ^2 с $k(k-1)/2$ степенями свободы. То есть число степеней свободы равно половине числа недиагональных элементов таблицы.

Например, мнения относительно наибольшей эффективности одного из четырех типов защиты у 100 опрошенных лиц поменялось после стажировки согласно табл. 6. Сумма частот в главной диагонали – это число людей, не изменивших мнения. Проверим симметрию относительно главной диагонали, выделенной серым цветом. Вычислим статистику:

$$\chi^2_{\text{сим}} = \frac{(10-4)^2}{10+4} + \frac{(16-12)^2}{16+12} + \frac{(1-6)^2}{1+6} +$$

$$+ \frac{(8-4)^2}{8+4} + \frac{(4-12)^2}{4+12} + \frac{(7-2)^2}{7+2} = 14,82. \qquad (6)$$

Достигнутая значимость составляет 0,022. Слишком маловероятно получить такое отклонение от симметрии. Следовательно, на уровне значимости 0,02 отклоняем гипотезу симметрии и приступаем к анализу изменений мнений. Значимая асимметрия относительно главной диагонали говорит о направленной деформации мнений в результате стажировки.

Таблица 6

Table 6

Деформация мнений об опасности атак

Changing attitudes about the dangers of attacks

	Тип атаки	После стажировки			
		Тип 1	Тип 2	Тип 3	Тип 4
До стажировки	Тип 1	7	10	16	1
	Тип 2	4	3	8	4
	Тип 3	12	4	4	7
	Тип 4	6	12	2	11

Результаты исследования сравним с информационным подходом к этой же задаче. Средняя информация в пользу H_1 против H_0 может быть найдена согласно [3] по формуле

$$2I(H_1 : H_0) = 2\sum_{i \neq j} n_{ij} \ln \frac{2n_{ij}}{n_{ij} + n_{ji}}. \quad (7)$$

Достигнутая значимость при использовании этой информационной статистики, которая в данном примере равна 14,144, составляет 0,028. Следовательно, гипотезу H_0 отклоняем.

Понятно, что вместо опасностей атак могли быть изменения других убеждений, симпатий после каких-то мероприятий. Вообще говоря, любая квадратная таблица сопряженности при анализе симметрии может дать нетривиальную информацию.

5. МЕРЫ СВЯЗИ НОМИНАТИВНЫХ ПРИЗНАКОВ

Если признаки оказались взаимосвязаны, то есть гипотеза об их независимости была проверена и отвергнута, возможна оценка силы доказанной связи. Ее хочется видеть в привычном интервале величин, например, от –1 до +1 с нулевым значением при отсутствии связи или от нуля до единицы, если отрицательные значения бессмысленны.

Для измерения силы связи на основе частотных таблиц предложены и используются в статистике около десятка формул [6], которые можно свести к трем основным группам: а) традиционные коэффициенты связи, использующие статистику χ^2; б) меры связи направленного типа; в) коэффициенты на основе информационных характеристик влияния признаков.

Известны многочисленные коэффициенты связи двух признаков, представленных таблицей сопряженности 2×2, т. е. оба признака представлены в двух категориях значений. Это коэффициент ассоциации Пирсона, коэффициент контингенции, частотный коэффициент детерминации, критерий Мак-Нимара и др. [6–8].

Рассмотрим **частотный коэффициент детерминации**, используемый для квадратных таблиц произвольного размера и тесно связанный с симметрией. Он вычисляется по формуле

$$\eta^2 = \frac{\sum_{i=1}^{k} n_{ii} - \sum_{i=1}^{k} n_{ii}^0}{\sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij} - \sum_{i=1}^{k} n_{ii}^0}, \tag{8}$$

где n_{ii}^0 – ожидаемые частоты на главной диагонали, вычисляемые по формуле (2). Значимость найденного коэффициента η^2 проверяем так же, как при проверке симметрии, предполагая асимптотическое распределение χ^2 с $k(k-1)/2$ степенями свободы.

Пример. Сотрудники организации оценивали в четырехбалльной шкале надежность режима секретности до и после неких модификаций. Изменение мнений представлено в табл. 7.

Таблица 7

Table 7

Пример изменения мнений о надежности защиты

Example of changing opinions about the reliability of protection

	Оценки после модификаций					
		2	**3**	**4**	**5**	Итого
Оценки до модификаций	**2**	3	3	5	2	13
	3	1	2	6	1	10
	4	6	6	2	1	15
	5	1	1	2	1	5
	Итого	11	12	15	5	43

Визуальное изучение таблицы сложно. Вычисление по формуле (8) дает $\eta^2 = -0{,}126$. Связь мнений «до» и «после» мала. Отрицательное значение говорит о том, что большинство респондентов всё же поменяли мнение о надежности. Однако асимптотическая значимость, равная 0,199, не позволяет признать эту асимметрию таблицы существенной. Это может быть случайным.

Рассмотрим **меры, использующие статистику** $\chi^2_{\text{эмп}}$ (2), позволяющие оценить связь двух признаков, измеренных в шкалах с различным числом категорий. То есть таблицы сопряженности признаков могут и не быть квадратными.

Коэффициент сопряженности Пирсона вычисляется по формуле

$$P = \sqrt{\frac{\chi^2_{\text{эмп}}}{\chi^2_{\text{эмп}} + n}}, \tag{9}$$

где $\chi^2_{\text{эмп}}$ находим по формуле (2). Минимальное значение P равно нулю, а максимальное не достигает единицы и зависит от размерности таблицы. Так что сопоставлять по этому коэффициенту можно только таблицы с одинаковым числом клеток. Коэффициент P можно улучшить [6, 1] делением на

$$P_{\max} = \sqrt{\frac{\min\{k,l\} - 1}{\min\{k,l\}}}, \tag{10}$$

и тогда его значение попадает в интервал [0,1].

Показатель Чупрова считается более объективным и вычисляется по формуле

$$T = \sqrt{\frac{\chi^2_{\text{эмп}}}{n\sqrt{(k-1)(l-1)}}}. \tag{11}$$

Показатель T достигает единицы, когда все частоты лежат на одной из самых длинных диагоналей таблицы. В квадратных таблицах ($k = l$) максимум T равен единице.

Для неквадратных таблиц часто используют **показатель Крамера**, который вычисляют по формуле

$$C = \sqrt{\frac{\chi^2_{\text{эмп}}}{\min\{k,l\}[(k-1)(l-1)]}}\,. \qquad (12)$$

Он всегда достигает единицы при полной связи и совпадает с T в квадратной таблице.

Для примера, показанного в табл. 8, значения этих мер следующие: $P = 0{,}419$; $T = 0{,}294$; $C = 0{,}326$.

Таблица 8

Table 8

Пример для мер на основе χ^2

Example for measures based on χ^2

Категории признака 1	Категории признака 2				Итоги
	Плохо	Удовлетв.	Хорошо	Отлично	$n_{i.}$
Низкая	19	12	9	2	42
Средняя	7	16	14	5	42
Высокая	4	11	26	12	53
$n_{.j}$	30	39	49	19	137

Достигнутая значимость при $\chi^2_{\text{эмп}} = 29{,}104$, найденная по (2), равна 0,00006, так что все эти меры, являющиеся функциями от χ^2, можно считать объективными. Эти меры всегда выражаются неотрицательными числами (заключение о знаке связи здесь лишено смысла), поэтому выяснение характера зависимости, ее специфических черт должно определяться непосредственно по таблице сопряженности.

Во многих ситуациях один из признаков явно зависит от другого. Скажем, потенциальная опасность, риск определяется типом вируса, но не наоборот. В таких случаях нужно использовать направленные меры.

Логично потребовать, чтобы мера связи таких признаков отражала бы возможность прогноза зависимого признака по значениям другого.

В табл. 9 показаны результаты проверки эффективности способа защиты от хакерских атак трех типов.

Таблица 9

Table 9

Пример для направленных мер

Example for targeted measures

Результат	Тип атаки			
	A	*B*	*C*	$n_{i.}$
Успех	13	16	7	36
Неудача	9	2	17	28
$n_{.j}$	22	18	24	$n_{..} = 64$

Число отраженных атак составляет 36 из 64, т. е. 56,25 %. Однако в 43,75 % случаев мы ошибемся, предсказывая успех. Если учесть тип атаки, тогда, предсказывая успех или неудачу по большинству экспериментальных результатов, ошибаемся в 9 + 2 + 7 = 18 случаев из 64 и процент ошибок предсказаний достигнет 28,13 %. Относительное уменьшение процента ошибок при учете типа атаки называют **показателем λ**. В этом примере получаем λ = (43,75 – 28,13) / 43,75 = 0,357.

Если учет типа атаки сведет процент ошибок к нулю, то λ станет равной единице и зависимость результата от типа атаки будет абсолютной, не случайной. Если учет типа атаки не улучшит прогноз результата, λ будет равна нулю, следовательно, зависимости мы не обнаружили.

Рассмотрим **показатель τ Гудмена – Крускала**. Если частоты результатов 36 и 28 взять с учетом процентов, то грубый прогноз даст 36 · 0,5625 + + 28 · 0,4375 = 32,53 угадываний из 64 и соответственно 31,48 ошибок. Это составляет 49,22 % от 64. Уточненный прогноз с учетом типа атаки и процентов успехов и неудач при каждом конкретном типе составит 13 · 0,5909 + 9 · 0,4091 + 16 · 0,8889 + 2 · 0,1111 + 7 · 0,2917 + 17 · 0,7083 = = 39,89 точных предсказаний и 64 – 39,89 = 24,11 ошибок, что даст 37,67 % от 64. Относительное уменьшение процента ошибок при учете типа атаки при таком подсчете получится равным τ = (49,22 – 37,67) / 49,22 = 0,235.

При проверке сопряженности таблицы по критерию (2) достигнутая значимость равна 0,001 и результаты не оставляют сомнений.

6. ИНФОРМАЦИОННЫЕ МЕРЫ СВЯЗИ

Рассмотрим меры связи между признаками (строками X и столбцами Y) таблицы в виде взаимной информации, т. е. уменьшения энтропии одного признака при известном другом.

Вернемся к табл. 1, где сведены частоты сопряжения категорий двух признаков. Если перейти к относительным частотам $p_{ij} = n_{ij} / n_{..}$, то оценки энтропии признаков можно получить по формуле Шеннона

$$H(X) = -\sum_{i=1}^{k} p_{i.} \log p_{i.} \; ; \tag{13}$$

$$H(Y) = -\sum_{j=1}^{l} p_{.j} \log p_{.j} \; . \tag{14}$$

Оценка энтропии исследуемой системы двух признаков (X, Y) равна

$$H(X,Y) = -\sum_{i=1}^{k} \sum_{j=1}^{l} p_{ij} \log p_{ij} \; . \tag{15}$$

Взаимная информация об одном признаке при известном состоянии другого будет равна уменьшению суммарной энтропии вследствие учета фактической зависимости признаков, проявляющейся в таблице:

$$I_{X \leftrightarrow Y} = H(X) + H(Y) - H(X,Y) \; . \tag{16}$$

В анализе используют информационные коэффициенты, получаемые путем следующей нормировки:

$$R_{X \leftrightarrow Y} = \frac{I_{X \leftrightarrow Y}}{H(X,Y)} \; , \tag{17}$$

а также направленные коэффициенты связи признаков, то есть качества прогноза одного признака по другому:

$$R_{X \to Y} = \frac{I_{X \leftrightarrow Y}}{H(Y)} \; ; \tag{18}$$

$$R_{Y \to X} = \frac{I_{X \leftrightarrow Y}}{H(X)} \; . \tag{19}$$

Допустим, в результате наблюдений эффективности способов защиты информации была составлена следующая таблица (табл. 10).

Таблица 10

Table 10

Пример для расчета информативности и дивергентности

Example for calculating of informativeness and divergence

		Эффективность защиты			Итоги
		Низкая (1)	Средняя (2)	Высокая (3)	$n_{i.}$
Способы обнаружения	a	24	7	7	38
	b	76	38	70	184
	c	69	32	82	183
	d	27	9	55	91
	$n_{.j}$	196	86	214	$n = 496$

Переходим к относительным частотам и получаем оценки: $H(X) = 1{,}7942$; $H(Y) = 1{,}4910$; $H(X, Y) = 3{,}2478$; $I_{X \leftrightarrow Y} = 0{,}0374$; $R_{Y \to X} = 0{,}0251$; $R_{X \to Y} = 0{,}0208$; $R_{X \leftrightarrow Y} = 0{,}0115$. Сопряженность признаков по критерию χ^2 достаточно высока, значимость достигает 0,00035. Однако их взаимная информация и коэффициенты связи очень малы, и их трудно интерпретировать в контексте анализа безопасности.

Для сравнительного анализа номинативных признаков в задачах классификации и принятия решений предлагается [3, 9] оценивать информативность признака как средневзвешенное количество информации, приходящееся на различные его категории. При этом категории другого признака воспринимаются как классы, принадлежность к которым отражает этот показатель информативности. Используя обозначения табл. 1, можно записать показатель информативности Шеннона [9–14] признака 1 для прогноза признака 2 как снятую неопределенность этой задачи классификации эффективности:

$$IS = 1 + \sum_{i=1}^{k} p_i \sum_{j=1}^{l} q_{ji} \log_k q_{ji} \ , \qquad (20)$$

где l – количество классов; k – количество категорий признака; q_{ji} – оценка условной вероятности попадания признака эффективности в j-й класс

(при условии, что используем *i*-й способ обнаружения), равная $q_{ji} = n_{ij} / n_{i}.$; p_i – оценка вероятности попадания признака в *i*-ю категорию, равная $p_i = n_i. / n.$.

Для табл. 10 получаем $IS = 0{,}2732$. Таким образом, информативность признака 1 (способ защиты) для предсказания признака 2 (эффективность) заметно больше аналогичного информационного коэффициента $R_{X \to Y} = 0{,}0208$. Но, главное, диапазон значений этого показателя информативности составляет [0, 1]. Эта нормировка позволяет сравнить и объективно выбрать наилучший признак из принципиально возможных типа: топология сети, протокол, программа.

Теория информации предлагает еще один инструмент для принятия решений – это расхождение (дивергенция).

Допустим, что распределение вероятностей некоторого признака *X* по *n* категориям согласно двум гипотезам H_j, $j = 1,\ 2$, таково: $p_{1j}, p_{2j}, \ldots, p_{kj}$. Следуя [3], можно найти среднюю информацию в пользу гипотезы H_1 против гипотезы H_2 при наблюдении *X* в условии справедливости H_1, равную

$$I_{KL}(1:2) = \sum_{i=1}^{k} p_{i1} \ln \frac{p_{i1}}{p_{i2}}. \tag{21}$$

Мы получили известную дивергенцию Кульбака – Лейблера [15], которую используют для оценки информационного выигрыша при замене распределения одного признака распределением другого, имеющую вид математического ожидания разности логарифмов вероятностей попадания в категории значений при H_1 и при H_2 при справедливости H_1.

В противоположном случае при справедливости H_2 средняя информация в пользу H_2 равна

$$I_{KL}(2:1) = \sum_{i=1}^{k} p_{i2} \ln \frac{p_{i2}}{p_{i1}} = -\sum_{k=1}^{k} p_{i2} \ln \frac{p_{i1}}{p_{i2}} \tag{22}$$

и отличается от $I_{KL}(1:2)$. То есть дивергенция Кульбака – Лейблера не является симметричной и не может служить оценкой различия распределений. Это направленная мера.

Например, для табл. 10 попарные сравнения распределений способов защиты при разных эффективностях дают такие дивергенции:

$I_{KL}(1:2) = 0,01772 \neq I_{KL}(2:1) = 0,016326;$

$I_{KL}(1:3) = 0,111845 \neq I_{KL}(3:1) = 0,093937;$

$I_{KL}(2:3) = 0,102129 \neq I_{KL}(3:2) = 0,113980.$

В работах [16, 17] предлагается симметричная мера в виде дивергенции Дженсена – Шеннона:

$$I_{JS}(1:2) = I_{JS}(2:1) = \frac{1}{2}\sum_{i=1}^{k} p_{i1} \ln \frac{2p_{i1}}{p_{i1}+p_{i2}} + \frac{1}{2}\sum_{i=1}^{k} p_{i2} \ln \frac{2p_{i2}}{p_{i1}+p_{i2}}. \qquad (23)$$

Для того же примера имеем $I_{JS}(1:2) = 0,004237$; $I_{JS}(1:3) = 0,024584$; $I_{JS}(1:3) = 0,026260$.

В статье [18] показано применение этой меры для распознавания аномальных изменений законов распределения информационных состояний ресурсов беспилотных транспортных средств, таких как канал связи, процессор, память. Предположительно, эти аномалии являются следствием атак.

Информационные подходы на этой основе позволяют избежать ошибок проверки гипотезы об отсутствии изменений и ориентированы на проверку набора из нескольких рабочих гипотез [19, 20]. В частности, показано, что статистика (2) для проверки гипотез уступает данному подходу. Затем несложно ранжировать эти гипотезы от лучших к худшим и масштабировать, чтобы использовать как вес правдоподобия каждой гипотезы.

Вместе с тем в работе [3] вводится другая симметричная оценка, которую можно интерпретировать как меру трудности различения двух гипотез H_1 и H_2 или двух классов (например, низкой и средней эффективности (табл. 10)) по законам распределения соответствующих им категорий признака, т. е. способов защиты:

$$J(1:2) = I_{KL}(1:2) + I_{KL}(2:1) = \sum_{i=1}^{k} (p_{i1} - p_{i2}) \ln \frac{p_{i1}}{p_{i2}}, \qquad (24)$$

где k – число категорий признака; p_{i1} и p_{i2} – оценки условных вероятностей низкой и средней эффективности при применении i-го способа защиты.

Другими словами, если мала дивергенция J на данной паре классов, то наш эксперимент не позволяет уверенно выбрать эффективный способ защиты.

Расчет для трех пар классов эффективности: низкая, средняя и высокая по табл.10 дает $J(1:2) = 0{,}03404$, $J(1:3) = 0{,}205782$ и $J(2:3) = 0{,}216109$. Можно заключить, что хуже всего по этому признаку различаются низкая и средняя эффективности защиты. Понятно, что этот результат явно зависит от наших предварительных правил классификации эффективности при заполнении таблицы, но теперь у нас появляется критерий для этих правил в виде максимума (23), либо мы в итоге убедимся, что исследуемые способы защиты фактически не отличаются по эффективности. Наконец, можно сравнивать различные признаки по показателям (21), (23) и (24) и обоснованно выбирать наиболее информативные.

ЗАКЛЮЧЕНИЕ

В статье показан подход к анализу ситуаций на основе наблюдений номинативных признаков. Примеры на основе методов проверки гипотез относительно связи признаков и способов оценки степени этой связи и информативности признаков ориентированы на проблематику защиты информации, в основном на стратегические, а не на оперативные задачи безопасности. Показано, что подход с позиций теории информации является сильным конкурентом классическим процедурам проверки гипотез, особенно в задачах классификации и принятия решений на ее основе.

Возможно, что на небольших таблицах сопряженности связи признаков могут быть легко замечены визуально. Однако в сложных случаях в задачах большого масштаба при сравнении более двух признаков без математического подхода, ограничиваясь интуицией и здравым смыслом, получить обоснованные выводы нереально.

Понятно, что некоторые постановки задач информационной безопасности в условиях номинативных признаков не сводятся к анализу таблиц сопряженности. Авторы будут признательны за конструктивные замечания и советы такого рода.

СПИСОК ЛИТЕРАТУРЫ

1. *Закс Л.* Статистическое оценивание. – М.: Статистика, 1976. – 599 с.

2. *Кендалл М., Стьюарт А.* Многомерный статистический анализ и временные ряды. – М.: Наука, 1976. – 736 с.

3. *Кульбак С.* Теория информации и статистика. – М.: Наука, 1967. – 408 с.

4. *Крутохвостов Д.С., Хиценко В.Е.* Парольная и непрерывная аутентификация по клавиатурному почерку средствами математической статистики // Вопросы кибербезопасности. – 2017. – № 5 (24). – С. 91–99.

5. *Рунион Р.* Справочник по непараметрической статистике. – М.: Финансы и статистика, 1982. – 200 с.

6. *Бююль А., Цёфель П.* SPSS: искусство обработки информации. – СПб.: Диасофт, 2002. – 602 с.

7. *Миркин Б.Г.* Анализ качественных признаков и структур. – М.: Статистика, 1980. – 320 с.

8. *Хиценко В.Е.* Математическая статистика для мониторинга информационной безопасности. Непараметрические методы статистики в примерах и задачах. – Saarbrücken: Lap Lambert Academic Publishing, 2013. – 208 с.

9. *Гублер Е.В.* Вычислительные методы анализа и распознавания патологических процессов. – Л.: Медицина, 1978. – 294 с.

10. *Колесникова С.И.* Методы анализа информативности разнотипных признаков // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. – 2009. – № 1 (6). – С. 69–80.

11. *Салахутдинова К.И., Лебедева И.С., Кривцова И.Е.* Подход к выбору информативного признака в задаче идентификации программного обеспечения // Научно-технический вестник информационных технологий, механики и оптики. – 2018. – Т. 18, № 2. – С. 278–285.

12. *Коржук В.М.* Модель и метод идентификации атак сетевого уровня на беспроводные сенсорные сети на основе поведенческого анализа: дис. … канд. техн. наук: 05.13.19. – СПб., 2019. – 206 с.

13. *Быкова В.В., Катаева А.В.* Методы и средства анализа информативности признаков при обработке медицинских данных // Программные продукты и системы. – 2016. – № 2. – С. 172–178.

14. Informativeness of genetic markers for inference of ancestry / N.A. Rosenberg, L.M. Li, R. Ward, J.K. Pritchard // American Journal of Human Genetics. – 2003. – Vol. 73 (6). – P. 1402–1422.

15. *Kullback S., Leibler R.A.* On information and sufficiency // Annals of Mathematical Statistics. – 1951. – Vol. 22 (1). – P. 79–86.

16. *Lin J.* Divergence measures based on the Shannon entropy // IEEE Transactions on Information Theory. – 1991. – Vol. 37, N 1. – P. 145–151.

17. *Nielsen F., Nock R.* Total Jensen divergences: definition, properties and clustering // 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – South Brisbane, QLD, Australia, 2015. – P. 2016–2020.

18. *Брюховецкий А.А.* Модель обнаружения аномальных данных на основе информационного критерия // Дневник науки. – 2021. – № 4. – URL: www.dnevniknauki.ru/images/publications/2021/4/technics/Bryukhovetskiy.pdf (дата обращения: 09.03.2022).

19. *Burnham K.P., Anderson D.R.* Kullback–Leibler information as a basis for strong inference in ecological studies // Wildlife Research. – 2001. – Vol. 28 (2). – P. 111–119.

20. *Do M.N.* Fast approximation of Kullback–Leibler distance for dependence trees and hidden Markov models // IEEE Signal Processing Letters. – 2003. – Vol. 10, N 4. – P. 115–118.

Хиценко Владимир Евгеньевич, кандидат технических наук, доцент кафедры защиты информации Новосибирского государственного технического университета. E-mail: xicenko@corp.nstu.ru

Федотов Никита Андреевич, магистрант кафедры вычислительной техники Новосибирского государственного технического университета. E-mail: nicoss.fd@yandex.ru

DOI: 10.17212/2782-2230-2022-1-61-84

Possibilities of analysis of nominative signs in tasks of information security *

V.E. Khitsenko[1], N.A. Fedotov[2]

[1] *Novosibirsk State Technical University, 20 Karl Marx Avenue, Novosibirsk, 630073, Russian Federation, candidate of technical sciences, associate Professor of the Department of Information Security. E-mail: xicenko@corp.nstu.ru*
[2] *Novosibirsk State Technical University, 20 Karl Marx Avenue, Novosibirsk, 630073, Russian Federation, computer science student. E-mail: nicoss.fd@yandex.ru*

Using various examples, the article demonstrates and discusses the possibilities of testing hypotheses and applying information measures to identify and assess the strength of the connection of nominative features in classification problems in the analysis of information security. The main type of presentation of the initial data in this scale is a contingency table of nominative features or an "object-feature" table, from which frequencies of coincidence of feature categories and a contingency table can be obtained. Using this table, it is easy to test the hypothesis of independence or homogeneity of features. An alternative approach to this analysis is considered based on the Kullback statistics, which is the average discriminating information in favor of the hypothesis of the dependence of features. In particular cases, the hypothesis of the symmetry of square tables is of practical interest, which can also be tested on the basis of information measures and criteria. An example of the processing of dichotomous data of the "yes-no" type according to the Cochran test is shown. The paper discusses ways to measure the strength of the connection of features. Illustrative examples of calculating measures based on chi-square statistics and directed measures are considered. The possibilities of various information characteristics are discussed in the form of a relative decrease in the entropy of one feature

* *Received 10 February 2022.*

with a known other, or in the form of a weighted average amount of information falling on different categories of a feature. These measures are useful for comparative analysis of nominative features in decision-making problems. Shannon's informativeness index, Kullback-Leibler divergence, and a measure of pairwise differentiation of protection efficiency classes according to the laws of distribution of the corresponding categories of a feature are used. The classical procedures for testing hypotheses and approaches based on information characteristics are consistently compared. The methods and examples considered in the work cover many urgent problems of information security associated with nominative features.

Key words: statistical methods, information protection, testing hypotheses about the conjugation of qualitative features, Cochran's and Kullback's criteria, testing the symmetry of tables, measures of the relationship of features, an informational approach to analyzing the relationship of features, Shannon's informativeness index, Kullback-Leibner and Jenssen-Shannon divergence

REFERENCES

1. Sachs L. *Statisticheskoe otsenivanie* [Statistical estimation]. Moscow, Statistika Publ., 1976. 599 p. (In Russian).

2. Kendall M., Stuart A. *Mnogomernyi statisticheskii analiz i vremennye ryady* [Multivariate statistical analysis and time series]. Moscow, Nauka Publ., 1976. 736 p. (In Russian).

3. Kullback S. *Teoriya informatsii i statistika* [Information theory and statistics]. Moscow, Nauka Publ., 1967. 408 p. (In Russian).

4. Krutohvostov D.S., Khitsenko V.E. Parol'naya i nepreryvnaya autentifikatsiya po klaviaturnomu pocherku sredstvami matematicheskoi statistiki [Password authentication and continuous authentication by keystroke dynamics using mathematical statistics]. *Voprosy kiberbezopasnosti = Cybersecurity Issues*, 2017, no. 5 (24), pp. 91–99.

5. Runyon R. *Spravochnik po neparametricheskoi statistike* [Nonparametric statistics]. Moscow, Finansy i statistika Publ., 1982. 200 p. (In Russian).

6. Bühl A., Zöfel P. *SPSS: iskusstvo obrabotki informatsii* [*SPSS:* The art of information processing]. St. Petersburg, Diasoft Publ., 2002. 602 p. (In Russian).

7. Mirkin B.G. *Analiz kachestvennykh priznakov i struktur* [Analysis of qualitative features and structures]. Moscow, Statistika Publ., 1980. 320 p.

8. Khitsenko V.E. *Matematicheskaya statistika dlya monitoringa informatsionnoi bezopasnosti. Neparametricheskie metody statistiki v primerakh i zadachakh* [Mathematical statistics for information security monitoring. Nonparametric methods of statistics in examples and problems]. Saarbrücken, Lap Lambert Academic Publishing, 2013. 208 p.

9. Gubler E.V. *Vychislitel'nye metody analiza i raspoznavaniya patologicheskikh protsessov* [Computational methods of analysis and recognition of pathological processes]. Leningrad, Meditsina Publ., 1978. 294 p.

10. Kolesnikova S.I. Metody analiza informativnosti raznotipnykh priznakov [Methods of analysis of different-type features informativity]. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naya tekhnika i informatika = Tomsk State University Journal of Control and Computer Science*, 2009, no. 1 (6), pp. 69–80.

11. Salakhutdinova K.I., Lebedeva I.S., Krivtsova I.E. Podkhod k vyboru informativnogo priznaka v zadache identifikatsii programmnogo obespecheniya [Informative feature selection in software identification task]. *Nauchno-tekhnicheskii vestnik informatsionnykh tekhnologii, mekhaniki i optiki = Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2018, vol. 18, no. 2, pp. 278–285.

12. Korzhuk V.M. *Model' i metod identifikatsii atak setevogo urovnya na besprovodnye sensornye seti na osnove povedencheskogo analiza*. Diss. kand. tekhn. nauk [Model and method of identification of network layer attacks on wireless sensor networks based on behavioral analysis. PhD eng. sci. diss.]. St. Petersburg, 2019. 206 p.

13. Bykova V.V., Kataeva A.V. Metody i sredstva analiza informativnosti priznakov pri obrabotke meditsinskikh dannykh [Methods and tools for analysing informative features when processing medical data]. *Programmnye produkty i sistemy = Software and Systems*, 2016, no. 2, pp. 172–178.

14. Rosenberg N.A., Li L.M., Ward R., Pritchard J.K. Informativeness of genetic markers for inference of ancestry. *American Journal of Human Genetics*, 2003, vol. 73 (6), pp. 1402–1422.

15. Kullback S., Leibler R.A. On information and sufficiency. *Annals of Mathematical Statistics*, 1951, vol. 22 (1), pp. 79–86.

16. Lin J. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 1991, vol. 37, no. 1, pp. 145–151.

17. Nielsen F., Nock R. Total Jensen divergences: definition, properties and clustering. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Australia, 2015, pp. 2016–2020.

18. Bryukhovetskiy A.A. Model' obnaruzheniya anomal'nykh dannykh na osnove informatsionnogo kriteriya [Anomalous data detection model based on information criterion]. *Dnevnik nauki*, 2021, no. 4. (In Russian). Available at: www.dnevniknauki.ru/images/publications/2021/4/technics/Bryukhovetskiy.pdf (accessed 09.03.2022).

19. Burnham K.P., Anderson D.R. Kullback-Leibler information as a basis for strong inference in ecological studies. *Wildlife Research*, 2001, vol. 28 (2), pp. 111–119.

20. Do M.N. Fast approximation of Kullback–Leibler distance for dependence trees and hidden Markov models. *IEEE Signal Processing Letters*, 2003, vol. 10, no. 4, pp. 115–118.

Для цитирования:

Хиценко В.Е., Федотов Н.А. Возможности анализа номинативных признаков в задачах информационной безопасности // Безопасность цифровых технологий. – 2022. – № 1 (104). – С. 61–84. – DOI: 10.17212/2782-2230-2022-1-61-84.

For citation:

Khitsenko V.E., Fedotov N.A. Vozmozhnosti analiza nominativnykh priznakov v zadachakh informatsionnoi bezopasnosti [Possibilities of analysis of nominative signs in tasks of information security]. *Bezopasnost' tsifrovykh tekhnologii = Digital Technology Security*, 2022, no. 1 (104), pp. 61–84. DOI: 10.17212/2782-2230-2022-1-61-84.

БЕЗОПАСНОСТЬ ЦИФРОВЫХ ТЕХНОЛОГИЙ. – 2022. – № 1 (104). – 85–98

МЕТОДЫ И СИСТЕМЫ ЗАЩИТЫ ИНФОРМАЦИИ, ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ

УДК 004.056 DOI: 10.17212/2782-2230-2022-1-85-98

ИССЛЕДОВАНИЕ ПРИМЕНЕНИЯ ГРАФОВ ЗНАНИЙ ДЛЯ ГЕНЕРАЦИИ ВОПРОСОВ АУДИТА ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ*

А.В. ЛЕМЯКИН[1], А.Б. АРХИПОВА[2]

[1] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, лаборант кафедры защиты информации. E-mail: a.v.lemyakin@gmail.com*
[2] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, доцент кафедры защиты информации. E-mail: arhipova@corp.nstu.ru*

В статье представлены подходы к исследованию вопросов применения графов знаний для генерации вопросов аудита информационной безопасности. Показана значимость человеческого фактора в аудите информационной безопасности как элемента социально значимой деятельности. В качестве решения формирования объективной базы исследования предложен метод автоматической генерации вопросов на основе графов знаний, способных выявлять набор картежей, представляющих собой субъект-объектные отношения. Проанализированы существующие модели и набор динамически формируемых метрик. В частности, субъективный метод пятибалльной шкалы Макото Нагао и машинный метод автоматической системы оценки на основе метода *N*-грамм. Проанализированы достоинства и недостатки методов исследования, представлена постановка задачи на дальнейшее исследование.

Ключевые слова: граф знаний, повышение осведомленности, политики безопасности, автоматическая генерация вопросов, семантический анализ, контроль знаний, тесты

ВВЕДЕНИЕ

Самый опасный нарушитель в любой компании – это внутренний нарушитель. Так, например, по статистике компании Infowatch, доля утечек по вине сотрудников за 2020 год составила 72 % [1]. Для решения этой проблемы призван целый ряд мер. Так, в статье [2] приводится пять основных областей: повышение культуры информационной безопасности (ИБ) организации

* Статья получена 10 февраля 2022 г.

и обучение персонала, ИТ-компетенция сотрудников, внедрение и соблюдение сотрудниками политик безопасности организации, управление рисками, менеджмент ИБ. Самыми эффективными были признаны повышение культуры ИБ организации, обучение персонала и менеджмент ИБ. Обучение персонала также подразумевает тестирование. Это две взаимосвязанные вещи. Однако обычный контроль осведомленности о политиках безопасности имеет значительный недостаток – ручной способ создания вопросов.

Решением данной проблемы могла бы стать автоматическая генерация вопросов на основе графов знаний, способных выявить из анализируемого текста так называемые триплеты – картежи, состоящие из трех элементов: объекта, субъекта и предиката (отношение между объектом и субъектом). При этом в идеале необходимо соблюсти универсальность генератора для любых политик безопасности, чтобы любая организация могла без лишних усилий автоматически сгенерировать тесты на контроль знаний.

1. АВТОМАТИЧЕСКАЯ ГЕНЕРАЦИЯ ВОПРОСОВ

Генерация вопросов является задачей синтеза текста наряду с задачей анализа текста. Анализ может быть представлен следующими этапами [3].

1. Считывание текстовой информации из входного файла.
2. Выделение отдельных членов предложения. Реализуется любым из множества доступных способов (например, выделять слова от пробела до пробела и добавлять в массив).
3. Определение принадлежности части речи для каждого члена предложения (например, с помощью библиотеки на языке C# – DeepMorphy).
4. Подготовка семантических моделей, представленных в виде графов, где узлами являются слова, а ребрами – их отношения.

Синтез текста (вопросов) может быть представлен комбинацией двух способов из нижеприведенных классификаций (по методам генерации и по структуре исходных данных).

1.1. КЛАССИФИКАЦИЯ СПОСОБОВ ГЕНЕРАЦИИ ВОПРОСОВ

По методам генерации вопросов [4].

1. Логические методы (LOGIC) – методы, основанные на представлении функций и моделей в виде заранее заданных формул, на основании которых задача генерации вопросов сводится к применению этих формул для преобразования исходных данных.

2. Методы на основе машинного обучения (ML) – методы, способные обучиться по прецедентам, т. е. сделать общие выводы на основе частных наблюдений.

3. Гибридные методы (HM) – методы, сочетающие в себе как логические методы, так и методы на основе машинного обучения.

По структуре исходных данных [4].

1. Использующие текстовые данные (TEXT) – методы, которые используют входные данные в виде человекочитаемого текста.

2. Использующие графовые данные (GRAPH) – методы, входные данные которых представляют собой знания, использующие графовое описание. Примером является формат представления знаний RDF (Resource Description Framework), в котором знания представлены в виде триплетов (субъект, предикат, объект), описывающих направленную связь субъекта с объектом.

3. Использующие гибридные данные (HD) – методы, для которых входные данные подаются в гибридном представлении описанных выше структур исходных данных. Примером такого представления может быть человек, читающий текст, дополненный онтологией в формате RDF, которая описывает предметную область текста.

В работе [5] для интеллектуальной обучающейся системы (ИОС) используется технология OSTIS (Open Semantic Technology for Intelligent Systems) и автоматическая генерация субъективных и объективных вопросов.

Генерация вопросов происходит логическими методами, которые реализуются аппаратом дискретной математики. Относительного этого подхода можно выделить следующие виды стратегии генерации вопросов [5].

1. На основе элементов:

- на основе ролевого отношения – то, которое задает роль элементов в рамках некоторого множества;
- на основе бинарного отношения – множество отношений на множестве M, являющихся подмножеством декартова произведения множества M на самого себя.

2. На основе различных отношений между классами:

- отношения «включение», когда некоторые классы в базе знаний содержат несколько подклассов. Пример: «Частным случаем бинарного дерева является: вариант 1, вариант 2, и т. д.»;
- отношения «разбиение», когда в результате разбиения множества получается множество попарно непересекающихся множеств, объединение которых есть исходное множество;
- отношения «строгое включение» – частный случай отношения включения.

3. На основе идентификаторов, когда некоторые сущности в базе знаний могут иметь несколько идентификаторов (синонимов). Пример: «Ориентированное множество также называется __?».

4. На основе аксиом.

5. На основе свойств отношений – использование свойств рефлексивности, симметричности и транзитивности, которым могут соответствовать некоторые отношения в базе знаний. Пример: «Если $A = B$, и при этом $B = C$, то равны ли A и C?».

6. На основе примеров изображений, когда в базе знаний хранятся некоторые понятия, отношения, теоремы и поясняющие их изображения.

7. Генерация субъективных вопросов – вопросы на определение понятий и доказательства теорем.

Вопросы, автоматически генерируемые с использованием перечисленных выше стратегий, могут содержать повторные и неправильные вопросы, поэтому для обеспечения качества генерируемых вопросов необходимо сначала сохранить эти автоматически генерируемые вопросы в базе знаний подсистемы автоматической генерации вопросов, а затем использовать ручные или автоматические подходы (сравнение подобия между вопросами) для фильтрации повторных и неправильных вопросов [5].

1.2. ОЦЕНКА КАЧЕСТВА СГЕНЕРИРОВАННЫХ ВОПРОСОВ

Существует два принципиальных способа оценки [6]: субъективный и машинный.

Первый может быть реализован с помощью пятибалльной шкалы Макото Нагао, в которой количество баллов означает следующее:

1 балл – смысл предложения понятен, и не возникает никаких вопросов; грамматика, словоупотребление и стиль соответствуют общей структуре текста и не требуют постредактирования;

2 балла – смысл предложения понятен, но возникают большие проблемы с грамматикой, словоупотреблением и стилем;

3 балла – общий смысл предложения понятен, но смысл некоторых его частей вызывает сомнение из-за неправильного грамматического строя;

4 балла – присутствуют ошибки словоупотребления и стилистики, требуется обращение к оригиналу;

5 баллов – в предложении имеется большое количество грамматических, словоупотребительных и стилистических ошибок, смысл предложения с трудом можно понять после внимательного изучения.

Второй способ может быть реализован с помощью автоматической системы оценки на основе метода *N*-грамм (вероятности появления цепочки букв *N*-го порядка (*N*-грамм) в анализируемых текстах). Он представлен таким метриками [7]: BLEU, NIST, METEOR.

BLUE (Bilingual Evaluation Understudy) – основная идея: «Чем ближе машинный текст к профессиональному человеческому тексту, тем лучше». Происходит вычисление точности выявленных *n*-грамм, так как, например, при переводе текста с одного языка на другой машинным способом оригинал часто оказывается более емким.

NIST (Национальный институт стандартов и технологий США) – в отличие от метода BLEU, помимо точности вычисляется еще и то, насколько информативным является конкретный *N*-грамм. Так, например, предлог будет иметь значительно меньший вес, чем последовательность из прилагательного и существительного.

METEOR (Metric for Evaluation of Translation with Explicit Ordering) использует функции сопоставления синонимов вместе с точным соответствием слов.

1.3. ПРИМЕРЫ ИСПОЛЬЗОВАНИЯ ГЕНЕРАЦИИ ВОПРОСОВ

В работе [8] приводится самостоятельно разработанное решение на основе уже существующего приложения linkeddata trivia. Результаты приведены в сравнительной табл. 1.

Таблица 1

Table 1

Сравнение решений

Comparison of solutions

	linkeddata trivia	Самостоятельная разработка
Способ построения вопросов	По случайному субъекту	По поиску субъекта; случайному субъекту, определенному по классу; случайному субъекту, определенному по тематике вопросов
Время на поиск субъектов	Около 23 с	< 1 с
Время на построение вопросов	–	Около 5 с

Окончание табл. 1

End of Tab. 1

	linkeddata trivia	Самостоятельная разработка
Полный перевод	Нет	Неполный (возможна ситуация, когда литерал на русском отсутствует в базе данных, и тогда он заменяется на английский, поэтому некоторые предикаты и субъекты могут быть на английском)
Примеры результатов	Что является частью военного конфликта операции «Радуга»?	Кто, или что, или какой автор Мона Лиза?
Особенности	–	Поиск реализован за счет Google Custom Search. Есть возможность поиска по географической карте
Источник исходных данных	Семантическая сеть DBPedia	

В статье [9] приводится модель Graph2Seq, которая улучшает современный показатель BLEU-4 с 11,57 до 29,40 и с 25,99 до 59,59 в тестах WQ (WebQuestions) и PQ (PathQuestions) соответственно.

1.4. ПРОБЛЕМЫ ГЕНЕРАЦИИ ВОПРОСОВ

Самые распространенные проблемы в области автоматической генерации вопросов [3].

1. Сложность моделирования семантики. Наличие в текстах синонимов, омонимов, неточностей, неопределенностей высказывания – всё это вызывает определенные проблемы при анализе текста.

2. Использование контекста. Зачастую человек воспринимает информацию лишь благодаря тому, что понимает контекст, в ходе которого эта информация была получена. Программа же контекста не понимает.

3. Трудности формирования сложных вопросов. Большинство решений позволяют генерировать только самые простые вопросы (вопросы на выбор, вопросы на заполнение пробелов и т. д.) [5].

Проблему контекста можно решить разными способами. Так, в статье [10] используется генерация последовательности вопросов на основе заранее известной последовательности ответов.

Для генерации вопросов выполняется взаимодействие с двумя графами:

1) об информации из отрывка текста; используется для лучшего захвата контекстных зависимостей;

2) об информации из ответов; используется, чтобы сделать сгенерированные вопросы более релевантными заданным ответам.

Пример исходного отрывка текста с отметками ответов:

(1) A small boy named [John]1 was at the park one day.

(2) He was [swinging]2 [on the swings]3 and [his friend]4 named [Tim]5 [played on the slide]6.

(3) John wanted to play on the slide now.

(4) He asked Tim [if he could play on the slide]7.

(5) Tim said [no]8, and he cried.

Таблица 2

Table 2

Последовательность вопросов на основе последовательности ответов

Sequence of questions based on sequence of answers

Сгенерированный вопрос	Ответ
Who was at the park?	John
What was he doing there?	swinging
On what?	on the wings
Who was he with?	his friend
Named?	Tim
What was he doing?	played on the side
What did John asked him?	if he could play on the slide
What did he say?	no

В статье [11] используется модель Refine Network (RefNet), которая пытается имитировать человеческий процесс создания вопросов, сначала создавая первоначальный черновик, а затем уточняя его. Модель обращает внимание как на исходный отрывок, так и на черновой вопрос и создает грамматически более правильный и полный вопрос. Реализуется это за счет двух декодеров: один формирует черновой вариант, второй формирует конечный вариант.

При этом второй декодер поддается улучшению по таким показателям, как беглость и возможность быть отвеченным (answerability) за счет использования дополнительной модели Reward-RefNet. Она проводит сравнение уточненного вопроса с результатами базового алгоритма (baseline) и вознаграждает систему за изменения, которые улучшают соответствующую метрику во время обучения.

Для оценки беглости используется балл BLEU. Для оценки возможности быть отвеченным используется балл, фиксирующий, содержит ли вопрос требуемые именованные сущности, важные слова, служебные слова, типы вопросов.

Пример исходного отрывка текста (жирным шрифтом выделен ответ, относительно которого генерируется вопрос): Liberated by **Napoleon's** army in 1806, Warsaw was made the capital of the newly created Duchy of Warsaw.

Таблица 3

Table 3

Результаты генерации вопросов для разных моделей

Results of generating questions for different models

Модель	Сгенерированный вопрос
Baseline	What was the capital of the newly duchy of Warsaw?
RefNet	Who liberated Warsaw in 1806?
Reward-RefNet	Whose army liberated Warsaw in 1806?

Проблема генерации сложных вопросов также поддается решению. Например, в статье [12] предлагается комплексное решение DQG (Deep Question Generation), состоящее из трех частей:

1) новый кодировщик графа, который включает механизм внимания в нейронную сеть Gated Graph (GGNN); используется для динамического моделирования взаимодействия между различными семантическими отношениями;

2) усовершенствование вложений отрывков текста на уровне слов и представлений семантического графа на уровне узлов; используется для получения унифицированных представлений отрывков с учетом семантики для декодирования вопросов;

3) введение вспомогательной задачи выбора контента, которая тренируется совместно с декодированием вопросов; используется для выбора подходящих контекстов в семантическом графе для формирования правильной цепочки рассуждений.

Поверхностный вопрос генерируется из одного предложения и апеллирует лишь к одному факту, который можно извлечь из такого предложения.

Таблица 4

Table 4

Пример генерации простого вопроса

Example of generating a simple question

Исходный текст	Кислород используется в клеточном дыхании и высвобождается в результате фотосинтеза, который использует энергию солнечного света для производства кислорода из воды
Вопрос	Какой жизненный процесс производит кислород в присутствии света?
Ответ	Фотосинтез

Сложный (глубокий) вопрос, требующий рассуждений и понимания текста в целом, формируется из нескольких непересекающихся релевантных предложений.

Таблица 5

Table 5

Пример генерации глубокого вопроса

Example of generating a deep question

Отрывок *A*	Международный аэропорт Паго-Паго, также известный как аэропорт Тафуна, представляет собой общественный аэропорт, расположенный в 7 милях (11,3 км) к юго-западу от центрального делового района Паго-Паго, в деревне и на равнинах Тафуна на острове Тутуила в Американском Самоа
Отрывок *B*	Аэропорт Хуна – это государственный общественный аэропорт, расположенный в одной морской миле (2 км) к юго-востоку от центрального делового района Хуна, Аляска
Вопрос	Находятся ли международный аэропорт Паго-Паго и аэропорт Хуны на американской территории?
Ответ	Да

ЗАКЛЮЧЕНИЕ

Среди существующих моделей генерации вопросов есть достаточно впечатляющие представители [10–12], которые способны устранить известные проблемы (поверхностные вопросы, невозможность учесть контекст). При этом есть возможность использовать открытый исходный код, например

модель DQG [12]. Также есть работы, которые могут послужить примером для проведения семантического анализа исходного документа [3, 8]. Для контроля же качества полученных результатов подойдут такие метрики, как BLUE, NIST и METEOR. Таким образом, задача автоматического создания тестов с приемлемым уровнем качества выглядит вполне реализуемой.

СПИСОК ЛИТЕРАТУРЫ

1. InfoWatch. Утечки информации ограниченного доступа: отчет за 9 месяцев 2020 г. – Экспертно-аналитический центр InfoWatch, 2020. – URL: https://d-russia.ru/wp-content/uploads/2020/12/infowatch_2020_9_monts_data_leak.pdf (дата обращения: 09.03.2022).

2. Information security awareness, a tool to mitigate information security risk: a literature review / O.S. Dada, E.A. Irunokhai, C.J. Shawulu, A.M. Nuhu, E.E. Daniel // Innovative Journal of Science. – 2021. – Vol. 3, N 3. – P. 29–54. – URL: https://www.researchgate.net/publication/353849520_Information_Security_Awareness_a_Tool_to_Mitigate_Information_Security_Risk_a_Literature_Review (accessed: 09.03.2022).

3. *Клыч А.А., Коршунов А.Г., Стативко Р.У.* Разработка подхода для генерирования вопросов // Актуальные проблемы теории и практики обучения физико-математическим и техническим дисциплинам в современном образовательном пространстве: IV Всероссийская (с международным участием) научно-практическая конференция. – Курск, 2020. – С. 367–373. – URL: https://elibrary.ru/item.asp?id=44845075 (дата обращения: 09.03.2022).

4. Автоматическая генерация вопросов на основе текстов и графов знаний / М.А. Белянова, Г.И. Ревунков, Г.И. Афанасьев, Ю.Е. Гапанюк // Динамика сложных систем – XXI век. – 2020. – Т. 14, № 4. – С. 55–64. – DOI: 10.18127/j19997493-202004-06.

5. *Вэньцзу Л.* Онтологический подход к автоматической генерации вопросов в интеллектуальных обучающих системах // Доклады БГУИР. – 2020. – Т. 18, № 5. – С. 44–52. – DOI: 10.35596/1729-7648-2020-18-5-44-52.

6. *Переходько И.В., Мячин Д.А.* Оценка качества компьютерного перевода // Вестник Оренбургского государственного университета. – 2017. – № 2 (202). – С. 92–96. – URL: https://elibrary.ru/item.asp?id=28904673 (дата обращения: 09.03.2022).

7. *Легостина М.С.* Метрики оценки качества машинного перевода // Инноватика-2019: сборник материалов XV Международной школы-конференции студентов, аспирантов и молодых ученых. – Томск, 2019. – URL: https://elibrary.ru/item.asp?id=39222633 (дата обращения: 09.03.2022).

8. *Андреичев М.Д., Ференец А.А.* Разработка программного комплекса генерации вопросов по заданным субъектам при помощи семантической сети // Электронные библиотеки. – 2019. – Т. 22, № 2. – С. 68–71. – DOI: 10.26907/1562-5419-2019-22-2-68-94.

9. *Chen Y., Wu L., Zaki M.J.* Toward subgraph guided knowledge graph question generation with graph neural networks // Journal of Latex Class Files. – 2015. – Vol. 14, N 8. – URL: https://arxiv.org/abs/2004.06015 (accessed: 09.03.2022).

10. *Chai Z., Wan X.* Learning to ask more: semi-autoregressive sequential question generation under dual-graph interaction // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. – Association for Computational Linguistics, 2020. – P. 225–237. – DOI: 10.18653/v1/2020.acl-main.21.

11. Let's ask again: refine network for automatic question generation / P. Nema, A.K. Mohankumar, M.M. Khapra, B. Ravindran, B.V. Srinivasan // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). – Hong Kong, China, 2019. – DOI: 10.18653/V1/D19-1326.

12. Semantic graphs for generating deep questions / L. Pan, Y. Xie, Y. Feng, T.-S. Chua, M.-Y. Kan // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. – Association for Computational Linguistics, 2020. – P. 1463–1475. – DOI: 10.18653/v1/2020.acl-main.135.

Лемякин Артём Викторович, лаборант кафедры защиты информации Новосибирского государственного технического университета. Область научных интересов – информационная безопасность. E-mail: a.v.lemyakin@gmail.com

Архипова Анастасия Борисовна, доцент кафедры защиты информации Новосибирского государственного технического университета. Основное направление научных исследований – математическое моделирование в информационной безопасности, оценка качества социально значимой деятельности. E-mail: arhipova@corp.nstu.ru

DOI: 10.17212/2782-2230-2022-1-85-98

Research of application of knowledge graphs to generate information security audit questions*

A.V. Lemyakin[1], A.B. Arkhipova[2]

[1] *Novosibirsk State Technical University, Karl Marx Avenue, 20, Novosibirsk, 630073, Russian Federation, laboratory assistant of the Department of Information Security. E-mail: a.v.lemyakin@gmail.com*
[2] *Novosibirsk State Technical University, Karl Marx Avenue, 20, Novosibirsk, 630073, Russian Federation, Associate Professor of the Department of Information Security. E-mail: arhipova@corp.nstu.ru*

The article presents approaches to research on the use of knowledge graphs to generate information security audit issues. The significance of the human factor in the audit of information security as an element of socially significant activity is shown. As a solution to the formation of an objective research base, a method of automatically generating questions based on knowledge graphs capable of identifying a set of cartoons that are subject-object relations is proposed. Existing models and a set of dynamically generated metrics were analyzed. In particular, the subjective method of the five-point Makoto Nagao scale and the machine method of the automatic evaluation system based on the N-gram method. Advantages and disadvantages of research methods are analyzed, task setting for further research is presented.

Keywords: knowledge graph, awareness raising, security policies, automatic question generation, semantic analysis, knowledge control, tests

REFERENCES

1. InfoWatch. *Utechki informatsii ogranichennogo dostupa: otchet za 9 mesyatsev 2020 g.* [Restricted Information Leaks: 9 Month 2020 Report]. Available at: https://d-russia.ru/wp-content/uploads/2020/12/infowatch_2020_9_monts_data_leak.pdf (accessed 09.03.2022).

2. Dada O.S., Irunokhai E.A., Shawulu C.J, Nuhu A.M., Daniel E.E. Information security awareness, a tool to mitigate information security risk: a literature review. *Innovative Journal of Science*, 2021, vol. 3, no. 3, pp. 29–54. Available at: https://www.researchgate.net/publication/353849520_Information_Security_Awareness_a_Tool_to_Mitigate_Information_Security_Risk_a_Literature_Review (accessed 09.03.2022).

3. Klych A.A., Korshunov A.G., Stativko R.U. [Development of an approach for generating questions]. *Aktual'nye problemy teorii i praktiki obucheniya fiziko-matematicheskim i tekhnicheskim distsiplinam v sovremennom obrazovatel'nom*

* *Received 10 February 2022.*

prostranstve [Actual problems of the theory and practice of teaching physical, mathematical and technical disciplines in the modern educational space]. IV All-Russian (with international participation) scientific and practical conference, Kursk, 2020, pp. 367–373. (In Russian). Available at: https://elibrary.ru/item.asp?id=44845075 (accessed 09.03.2022).

4. Belyanova M.A., Revunkov G.I., Afanasyev G.I., Gapanyuk Yu.E. Avtomaticheskaya generatsiya voprosov na osnove tekstov i grafov znanii [Automatic question generation based on texts and knowledge graphs]. *Dinamika slozhnykh sistem – XXI vek = Dynamics of Complex Systems – XXI Century*, 2020, vol. 14, no. 4, pp. 55–64. DOI: 10.18127/j19997493-202004-06.

5. Wenzu L. Ontologicheskii podkhod k avtomaticheskoi generatsii voprosov v intellektual'nykh obuchayushchikh sistemakh [Ontological approach to automatic generation of questions in intelligent learning systems]. *Doklady Belorusskogo gosudarstvennogo universiteta informatiki i radioelektroniki = Doklady BGUIR*, 2020, vol. 18, no. 5, pp. 44–52. DOI: 10.35596/1729-7648-2020-18-5-44-52.

6. Perekhodko I.V., Myachin D.A. Otsenka kachestva komp'yuternogo perevoda [Quality evaluation of computer translation]. *Vestnik Orenburgskogo gosudarstvennogo universiteta = Vestnik of the Orenburg State University*, 2017, no. 2 (202), pp. 92–96. Available at: https://elibrary.ru/item.asp?id=28904673 (accessed 09.03.2022).

7. Legostina M.S. [Metrics of quality assessment of machine translate]. *Innovatika-2019* [Innovation-2019]. Collection of materials of the XV International School-Conference of Students, Graduate Students and Young Scientists, Tomsk, 2019. (In Russian). Available at: https://elibrary.ru/item.asp?id=39222633 (accessed 09.03.2022).

8. Andreichev M.D., Ferenets A.A. Razrabotka programmnogo kompleksa generatsii voprosov po zadannym sub"ektam pri pomoshchi semanticheskoi seti [Development of a program complex for generating questions on given subjects using a semantic network]. *Elektronnye biblioteki = Russian Digital Libraries Journal*, 2019, vol. 22, no. 2, pp. 68–94. DOI: 10.26907/1562-5419-2019-22-2-68-94.

9. Chen Y., Wu L., Zaki M.J. Toward subgraph guided knowledge graph question generation with graph neural networks. *Journal of Latex Class Files*, 2015, vol. 14, no. 8. – Available at: https://arxiv.org/abs/2004.06015 (accessed 09.03.2022).

10. Chai Z., Wan X. Learning to ask more: semi-autoregressive sequential question generation under dual-graph interaction. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020, pp. 225–237. DOI: 10.18653/v1/2020.acl-main.21.

11. Nema P., Mohankumar A.K., Khapra M.M., Ravindran B., Srinivasan B.V. Let’s ask again: refine network for automatic question generation. *Proceedings of*

the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 2019. DOI: 10.18653/V1/D19-1326.

12. Pan L., Xie Y., Feng Y., Chua T.-S., Kan M.-Y. Semantic graphs for generating deep questions. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020, pp. 1463–1475. DOI: 10.18653/v1/2020.acl-main.135.

Для цитирования:

Лемякин А.В., Архипова А.Б. Исследование применения графов знаний для генерации вопросов аудита информационной безопасности // Безопасность цифровых технологий. – 2022. – № 1 (104). – С. 85–98. – DOI: 10.17212/2782-2230-2022-1-85-98.

For citation:

Lemyakin A.V., Arkhipova A.B. Issledovanie primeneniya grafov znanii dlya generatsii voprosov audita informatsionnoi bezopasnosti [Research of application of knowledge graphs to generate information security audit questions]. *Bezopasnost' tsifrovykh tekhnologii = Digital Technology Security*, 2022, no. 1 (104), pp. 85–98. DOI: 10.17212/2782-2230-2022-1-85-98.

ПРАВИЛА ДЛЯ АВТОРОВ

УСЛОВИЯ ПРИЕМА СТАТЕЙ

Перед отправкой рукописи в редакцию авторам обязательно требуется проверить свою статью с помощью системы «Антиплагиат». Принятый редакционной коллегией уровень оригинальности статей должен составлять не менее 85 %.

Все статьи и сопровождающие их материалы в журнал подаются через сайт журнала в электронном виде после регистрации всех авторов статьи. Регистрация обязывает каждого автора иметь международный идентификационный номер ORCID. Иные варианты подачи материалов не рассматриваются.

Автор (один из соавторов) в своем личном кабинете выбирает в меню пункт «Подать статью» и вводит все необходимые данные. Своих соавторов при этом он выбирает из списка зарегистрированных пользователей.

Рукопись статьи готовится в соответствии с правилами оформления в редакторе MS Word и прикрепляется в формате *.doc, *.docx.

Сканированные лицензионный договор с подписями авторов и экспертное заключение (цветной режим сканирования, разрешение не менее 600 dpi) необходимо также прикрепить на сайте журнала в разделе «Подать статью» в формате *.pdf, *.jpg, *.jpeg.

По окончании всех работ обязательно нажать кнопку «Отправить в редакцию».

В редакцию журнала предоставляются следующие материалы.

1. **Статья**, подготовленная в соответствии с правилами оформления, – печатная версия, 2 экземпляра, подписанных авторами.

2. **Контактная информация** (телефоны рабочий и сотовый, адреса электронной почты, место работы, адрес места работы, должность, ученая степень, ученое звание автора) – печатная версия, 2 экземпляра.

3. **Описание статьи для базы данных «Российский индекс научного цитирования (РИНЦ)»**, подготовленное в соответствии с правилами оформления, – печатная версия, один экземпляр.

4. **Лицензионный договор**, заполненный и подписанный.

5. **Электронная версия статьи, контактной информации, описания статьи для базы данных РИНЦ, сканированный лицензионный договор и экспертное заключение о возможности опубликования (в отдельных файлах на адрес редакции)**.

6. **Согласие на публикацию, обработку и распространение персональных данных авторов статей**.

7. **Экспертное заключение о возможности опубликования**.

Редакцией рассматриваются только те материалы авторов, которые полностью соответствуют вышеобозначенным требованиям. Неполный пакет материалов редакцией не рассматривается.

Подготовленные материалы направляются на почтовый адрес редакции: 630073, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет (НГТУ), корп. 7, ком. 606, в редакцию журнала «Безопасность цифровых технологий».

Все рукописи рецензируются, по результатам рецензирования редколлегия принимает решение о целесообразности опубликования материалов.

ПРАВИЛА ОФОРМЛЕНИЯ

При подготовке документов для отправки в редакцию журнала авторам рекомендуется внимательно прочитать правила и посмотреть примеры оформления статей и всех необходимых сопутствующих документов. Редакция рассматривает статьи, подготовленные как на русском, так и на английском языке. Для опубликования статьи на английском языке необходимо дополнительно предоставить ее русскоязычный вариант, оформленный по правилам журнала (кроме зарубежных авторов).

Перед отправкой рукописи в редакцию авторы должны проверить свою статью с помощью системы «Антиплагиат». Принятый редакционной коллегией уровень оригинальности статей – не менее 85 %.

Чтобы статья была направлена на рецензирование, необходимо подготовить следующее:

1) **статью** в соответствии с правилами оформления (объем от 7 до 30 страниц);

2) **контактную информацию** в одном файле предоставить по каждому автору: ФИО полностью, ученая степень, ученое звание, должность, место работы, адрес места работы, телефон рабочий и мобильный, адрес электронной почты;

3) **описание статьи для базы данных «Российский индекс научного цитирования (РИНЦ)»**;

4) подписанный авторами **лицензионный договор** заполнить;

5) **экспертное заключение** о возможности опубликования, принятое в вашей организации;

6) согласие на публикацию, обработку и распространение персональных данных авторов статей;

7) авторы, не являющиеся сотрудниками НГТУ, предоставляют **сопроводительное письмо** на имя проректора по научной работе НГТУ (ссылка на

страницу сайта НГТУ). Письмо нужно подготовить на бланке организации с подписью и печатью руководителя.

ОСНОВНЫЕ РАЗДЕЛЫ ЖУРНАЛА

Автоматизация и управление технологическими процессами и производствами.

Управление в социальных и экономических системах.

Методы и системы защиты информации, информационная безопасность.

RULES FOR AUTHORS

CONDITIONS FOR ACCEPTANCE OF ARTICLES

Before sending the manuscript to the editorial office, authors must check their article using the Antiplagiarism system. The level of originality of articles adopted by the Editorial Board should be at least 85 %.

All articles and their accompanying materials are submitted to the magazine through the magazine's website in electronic form after registration of all the authors of the article. Registration obliges each author to have an international ORCID. No other material supply options are considered.

The author (one of the co-authors) in his personal account selects the item "Submit article" in the menu and enters all the necessary data. At the same time, he selects his co-authors from the list of registered users.

The manuscript of the article is prepared in accordance with the design rules in the MS Word editor and attached in the format * .doc, * .docx.

Scanned license agreement with signatures of authors and expert opinion (color mode scanning, resolution not less than 600 dpi) can also be attached on the website of the magazine in the section "Submit article" in the format * .pdf, * .jpg, * .jpeg.

At the end of all works, be sure to click the "Send to Design" button.

The following materials are provided to the journal editor:

1. **The article**, prepared in accordance with the rules of design, is a private version, 2 copies signed by the authors.
2. **Contact information** (working and cellular phones, e-mail addresses, place of work, address of the place of work, position, scientific degree, academic title of the author) – printed version, 2 copies.
3. **The description of the article** for the database "**Russian Scientific Citation Index (RSCI),**" prepared in accordance with the rules of form-making, is a printed version, one copy.
4. **License agreement** completed and signed.
5. **Electronic version of the article**, contact information, description of the article for the RSCI database, scanned license agreement and expert opinion on the possibility of publication (in separate files to the editorial address).
6. **Consent to the publication, processing and dissemination of the personal data** of the authors of the articles.
7. **Expert opinion** on the possibility of publication.

The editors consider only those materials of the authors that fully meet the above requirements. Incomplete package of materials is not considered by the revision.

The prepared materials are sent to the postal address of the editorial office: 630073, Novosibirsk, Karl Marx Avenue, 20, Novosibirsk State Technical University (NSTU), building 7, office 606, to the editors of the journal "Digital Technology Security."

All manuscripts were reviewed, and according to the results of the review, the editorial board decided on the appropriateness of publishing the materials.

FORMATTING RULES

When preparing documents for submission to the journal editor, authors are advised to carefully read the rules and see examples of the design of articles and all necessary related documents. The Drafting Committee considered articles prepared in both Russian and English. To publish the article in English, it is necessary to additionally provide its Russian-language version, drawn up according to the rules of the magazine (except for foreign authors).

Before sending the manuscript to the editorial board, authors should check their article using the Antiplagiarism system. The level of originality of articles adopted by the Editorial Committee was at least 85 %.

For the article to be aimed at peer review, you need to prepare the following:

1) **the article** in accordance with the rules of design (volume from 7 to 30 pages);

2) **provide contact information** in one file for each author: full name, degree, academic title of the author, position, place of work, address of the place of work, telephone number of the worker and mobile, e-mail address;

3) **description of the article** for the database "Russian Scientific Citation Index (RSCI)";

4) fill out the **license agreement**, the form of the license agreement must be signed only by the authors (it is also available to the authors in the personal office), if there are several authors, then it is necessary to add fields on all authors and sign each of them;

5) **expert opinion** on the possibility of publication, adopted in your organization;

6) consent to the publication, processing and dissemination of the personal data of the authors of the articles;

7) authors who are not employees of the NSTU provide a **companion letter** addressed to the vice-rector for scientific work of the NSTU (link to the page of the NSTU website). The letter should be prepared on the form of the organization with the signature and seal of the manager.

JOURNAL SECTION

Automation and control of technological processes and productions.
Governance in social and economic systems.
Methods and systems of information protection, information security.