ISSN 2713-3192
DOI 10.15622/ia.2021.20.4
http://ia.spcras.ru

ТОМ 20 № 4

ИНФОРМАТИКА И АВТОМАТИЗАЦИЯ

INFORMATICS AND AUTOMATION

СПб ФИЦ РАН

Санкт-Петербург
2021

INFORMATICS AND AUTOMATION

Volume 20 № 4, 2021

Scientific and educational journal primarily specialized in computer science, automation, robotics, applied mathematics, interdisciplinary research
Founded in 2002

Founder and Publisher
St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS)

Editor: I. O. Novikova
Interpreter: E.N. Mesheryakova
Art editor: N.A. Dormidontova

Editorial office address
14-th line V.O., 39, SPIIRAS, St. Petersburg, 199178, Russia,
e-mail: ia@spcras.ru, web: http://ia.spcras.ru
The journal is indexed in Scopus
The journal is published under the scientific-methodological supervision of Department for Nanotechnology and Information Technologies of the Russian Academy of Sciences

ИНФОРМАТИКА И АВТОМАТИЗАЦИЯ

Том 20 № 4, 2021

Научный, научно-образовательный журнал с базовой специализацией
в области информатики, автоматизации, робототехники, прикладной математики
и междисциплинарных исследований.
Журнал основан в 2002 году

Учредитель и издатель

Федеральное государственное бюджетное учреждение науки
«Санкт-Петербургский Федеральный исследовательский центр Российской академии наук»
(СПб ФИЦ РАН)

Адрес редакции

199178, г. Санкт-Петербург, 14-я линия В.О., д. 39
e-mail: ia@spcras.ru, сайт: http://ia.spcras.ru

Журнал индексируется в международной базе данных Scopus

Журнал входит в «Перечень ведущих рецензируемых научных журналов и изданий,
в которых должны быть опубликованы основные научные результаты диссертации
на соискание ученой степени доктора и кандидата наук»

Журнал выпускается при научно-методическом руководстве Отделения нанотехнологий
и информационных технологий Российской академии наук

CONTENTS

СОДЕРЖАНИЕ

УДК 004.056 **DOI** 10.15622/ia.20.4.1

И.В. Котенко, И.Б. Саенко, А.А. Браницкий,
И.Б. Паращук, Д.А. Гайфулина

ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА АНАЛИТИЧЕСКОЙ ОБРАБОТКИ ЦИФРОВОГО СЕТЕВОГО КОНТЕНТА ДЛЯ ЕГО ЗАЩИТЫ ОТ НЕЖЕЛАТЕЛЬНОЙ ИНФОРМАЦИИ

Котенко И.В., Саенко И.Б, Браницкий А.А., Паращук И.Б., Гайфулина Д.А. **Интеллектуальная система аналитической обработки цифрового сетевого контента для его защиты от нежелательной информации.**

Аннотация. В настоящее время Интернет и социальные сети как среда распространения цифрового сетевого контента становятся одной их важнейших угроз персональной, общественной и государственной информационной безопасности. Возникает необходимость защиты личности, общества и государства от нежелательной информации. В научно-методическом плане проблема защиты от нежелательной информации имеет крайне небольшое количество решений. Этим определяется актуальность представленных в статье результатов, направленных на разработку интеллектуальной системы аналитической обработки цифрового сетевого контента для защиты от нежелательной информации. В статье рассматриваются концептуальные основы построения такой системы, раскрывающие содержание понятия нежелательной информации и представляющие общую архитектуру системы. Приводятся модели и алгоритмы функционирования наиболее характерных компонентов системы, таких как компонент распределенного сканирования сети, компонент многоаспектной классификации сетевых информационных объектов, компонент устранения неполноты и противоречивости и компонент принятия решений. Представлены результаты реализации и экспериментальной оценки системных компонентов, которые продемонстрировали способность системы отвечать предъявляемым требованиям по полноте и точности обнаружения и противодействию нежелательной информации в условиях ее неполноты и противоречивости.

Ключевые слова: интеллектуальная система, цифровой сетевой контент, нежелательная информация, классификация, нечеткие знания, принятие решений.

1. Введение. Стремительное внедрение глобальной сети Интернет и построенных на ее основе социальных сетей в государственную, производственно-экономическую и социально-культурную сферы современного общества является мощным стимулом их дальнейшего развития. В то же время Интернет и социальные сети становятся одной их важнейших угроз персональной, общественной и государственной информационной безопасности. Возникает необходимость защиты личности, общества и государства от нежелательной информации, которая распространяется через глобальные компьютерные сети и способна нанести вред здоровью граждан или мотивировать их к противоправному поведению. В ведущих странах мира, включая Россию, защита от нежелательной сетевой информации регулируется национальным законодательством. Сайты с нежелательным контентом блокируются и заносятся в черные списки. Однако обнаружение нежелательных сайтов и

формирование черных списков осуществляется, как правило, в ручном режиме, а экспертное суждение о принадлежности информации к той или иной категории всегда является субъективным. По этой причине оно может быть недостаточно полным и/или ошибочным, что требует добавления в процесс выявления и противодействия нежелательной информации методов устранения ее неполноты и противоречивости.. Кроме того, при ручном режиме анализа Интернет-контента достаточно сложно обеспечить выполнение требований по своевременности реагирования на появление новых информационных объектов и изменение содержимого существующих ресурсов.

В научно-методическом плане проблема защиты от нежелательной информации имеет небольшое количество научно-технических решений. Несмотря на то, что за последние годы появились методики и реализации отдельных компонентов такого рода систем защиты, они или находятся на начальной стадии разработки и внедрения, или не реализуют полного спектра предполагаемых возможностей [1-3].

Целью настоящей статьи является изложение результатов исследований, посвященных построению и функционированию перспективной интеллектуальной системы, предназначенной для аналитической обработки цифрового сетевого контента в интересах защиты от нежелательной информации. Для выявления и противодействия нежелательной информации в данной системе используются методы машинного обучения и методы обработки нечетких данных. С целью удовлетворения этих требований в системе предложен и реализуется ряд специфических процессов обработки данных, которые также выделяют разработанную систему от других подобных систем. К числу этих процессов следует отнести: одновременный анализ большого количества источников данных о смысловом наполнении информационного объекта; многоуровневая классификация цифровых информационных объектов; использование методов обработки неполной и противоречивой информации; применение различных способов противодействия нежелательной информации в зависимости от целевой аудитории и некоторые другие.

К числу новых результатов исследований, которые освещает статья, относятся: (1) концептуальные основы построения и функционирования предлагаемой интеллектуальной системы; (2) модели и методы работы основных компонентов системы; (3) ключевые аспекты реализации и результаты экспериментальной оценки использования системы для решения возлагаемых на нее задач. Этим определяется дальнейшая структура статьи. В разделе 2 приводятся результаты анализа релевантных работ. Раздел 3 содержит концептуальные основы, раскрывающие понятие нежелательной информации, а также общую структуру целевой

системы. Реализация и экспериментальная оценка системы обсуждаются в разделе 4. Раздел 5 содержит заключительные выводы и направления дальнейших исследований.

2. Состояние исследований. Несмотря на то, что в последние годы появились методы и реализации отдельных компонентов такого рода систем защиты, они либо находятся на начальной стадии разработки и внедрения, либо не реализуют весь спектр ожидаемых возможностей. Так, в [4-9] рассматриваются некоторые механизмы обнаружения и противодействия вредоносной информации в сетевых информационных объектах. В этих документах излагаются решения для определения надежных оценок цифрового сетевого контента. Рассмотренные в них механизмы основаны на методах классификации информации, методах интеллектуальной обработки данных и фильтрации спама. Однако эти механизмы не ориентированы на работу в условиях семантической неопределенности информационного содержания.

В работах [10-12] рассматриваются различные методы анализа социальных сетей для обнаружения и выбора мер противодействия вредоносной информации. В [10] для обнаружения вредоносной информации используются алгоритмы поиска по описанию события, идентификации пользователей различных сетей и поиска по группам пользователей. Методы количественной и качественной оценки информационных воздействий в социальных сетях, основанные на табличных и графических инструментах для представления метрик и расчета метрик, обсуждаются в [11]. В [12] рассмотрены подходы к определению демографических характеристик пользователей социальных сетей. Однако, поскольку помимо социальных сетей существуют и другие источники нежелательной информации, эти подходы нельзя считать универсальными.

На наш взгляд, в наибольшей степени для обнаружения и противодействия нежелательной информации подходят методы анализа трафика на основе классификации веб-страниц. Эти методы могут быть основаны на контент-анализе внутренних свойств веб-страниц [13]. Бинарный классификатор, основанный на выявлении групп внутренних свойств HTML-документов, используется для обучения систем классификации веб-страниц в работах [14, 15]. В [16] показано, что обучение классификаторов обнаружению и противодействию нежелательной информации может быть реализовано на основе комбинации значимых функций веб-страниц. Однако методы, представленные в [13-16], не ориентированы на анализ содержания веб-страниц, то есть веб-контента.

В ряде работ обнаружение и противодействие нежелательной информации реализуется с использованием алгоритмов классификации тем веб-контента [17, 18]. В этом случае поиск вредоносной информации осуществляется по URL-адресам. Однако этот метод уменьшает спектр характеристик нежелательной информации, которые необходимо анализировать, и, соответственно, уменьшает диапазон контрмер. Некоторое время был популярен подход, основанный на анализе ссылок в веб-контенте. Такой анализ позволял реализовать иерархическую и объединенную классификацию веб-контента [19, 20]. Для классификации использовались модели на основе метода SVM. Для классификации веб-сайтов в [21] предложен алгоритм Link Information Categorization (LIC), который основан на методе классификации kNN. Классификация страниц с помощью алгоритма kNN также была исследована в [22], где различным терминам и тегам присваиваются соответствующие веса. В [23] выполняется классификация веб-сайтов с помощью анализа существительных, извлекаемых из веб-страниц. В качестве метода классификации используется Decision Tree. В [24] рассмотрен метод, заключающийся в поиске и извлечении значимого текста из тегов с последующим применением классификатора Naïve Bayes к полученным выборкам. Такой же подход в сочетании с методами противодействия вредоносной информации упоминается в [25, 26].

Выявление и противодействие нежелательной информации в реальных условиях, то есть когда обработка и оценка свойств нежелательной информации осуществляется в условиях неполноты и неопределенности, требует использования подходов, основанных на методах, моделях и алгоритмах устранения неопределенности и неполноты. Например, обработка неопределенной информации различного типа и поддержка принятия решений обычно реализуются с использованием искусственных нейронных сетей [27-30], нечетких множеств [31, 32] и нейронечетких сетей [33]. Применение этих методов для обнаружения и противодействия нежелательной информации является довольно сложной задачей. Однако преимущества этих методов заключаются в том, что они позволяют выбирать меры противодействия вредоносной информации на основе оценки семантического содержания информационных объектов в условиях неполноты и неопределенности. Эти методы также будут рассмотрены в статье.

3. Концептуальные основы интеллектуальной системы аналитической обработки цифрового сетевого контента. В настоящем разделе рассматриваются понятие нежелательной информации и общая архитектура предлагаемой системы.

Нежелательная информация воспринимается зачастую как элемент информационного воздействия. Информационный эффект *R* от информационного воздействия трактуется как основной поражающий фактор информационной войны. Он представляет собой воздействие информационным потоком на информационную систему как на объект атаки. Объектом атаки может являться отдельный человек, коллектив людей (некоторая организация) и даже государство в целом. Естественно, что эффект может быть как отрицательный, так и положительный. Однако воздействие с положительным эффектом мы рассматривать не будем. Поэтому будем считать, что цель такого воздействия заключается в достижении негативных структурных и/или функциональных изменений системы за счет приема и обработки этой информации. Формально информационный эффект определяется следующим образом:

$$R = IE\ (IO), \tag{1}$$

где *IE* — функция, определяющая некоторое информационное воздействие, *IO* — информационный объект, *R* — результат воздействия.

Информационный объект (ИО) *IO* есть логически цельный блок информации, представленный в определенной фиксированной форме, который создан и используется в информационной деятельности человека. Формально связь этого понятия с другими понятиями представляется следующим образом: $IO \in I$, то есть ИО является элементом множества всей анализируемой информации *I*.

Использование понятия ИО позволяет предложить другой вариант определения нежелательной информации, основывающийся на анализе информационных признаков ИО. Обозначим всю информация в сети Интернет как *Int*. Положим, что множество *Int* содержит опасную информацию *RI* (*Risky Information*) и безопасную информацию *SI* (*Safe Information*). Между этими понятиями справедливо следующее равенство:

$$Int = RI \cup SI\ . \tag{2}$$

Нежелательная информация (Inappropriate Information, II) есть отдельный ИО и/или совокупность объектов в сети Интернет, содержащих признаки, попадающие под категории ненужности, негодности. Наиболее ярким примером здесь являются ИО, фильтруемые системой

родительского контроля. Кроме того, будем относить к категории нежелательной информации также сомнительную и вредоносную информацию, упоминания которой иногда встречаются в литературе.

Сомнительная информация (*Dubious Information, DI*) есть отдельный ИО и/или совокупность объектов в сети Интернет, содержащие признаки, попадающие под категории опасности. Например, фишинговый сайт, недоверенный ресурс, ресурс с низкой репутацией. Объект, содержащий ложную (фейковую) информацию или дезинформацию, также относится к данному типу.

Вредоносная информация (Harmful information, HI) есть отдельный ИО и/или совокупность объектов в сети Интернет, содержащие признаки, по которым информация запрещена к распространению. Например, под эту категорию попадает информация, включенная в федеральный список экстремистских материалов, конфиденциальная информация, персональные данные и т.д.

Используя введенные обозначения для различных типов информации, можно сформировать между ними следующие соотношения:

$$II \subseteq RI\ ,\ \left(DI \cup HI\right) \subseteq II\ . \qquad (3)$$

Следует отметить, что в общем случае пересечение множеств *DI* и *HI* не является пустым множеством, то есть один и тот же ИО может быть отнесен как к *DI*, так и к *HI*.

Общая архитектура системы. Предлагаемая интеллектуальная система аналитической обработки цифрового сетевого контента (ИСАОЦСК) имеет общую архитектуру, показанную на рисунке 1. Архитектура содержит три уровня:

1) сбора и предварительной обработки данных о безопасности сетевых ИО;

2) оценивания смыслового содержания ИО;

3) выработки мер противодействия выявленной в ИО нежелательной информации.

Исходными данными для такой системы являются информационные объекты сети Интернет и социальных сетей. Результаты, полученные с помощью ИСАОЦСК, используются пользователями (администраторами безопасности), отвечающими за защиту от нежелательной информации. Потребителями результатов функционирования ИСАОЦСК являются регуляторы телекоммуникационного сектора.

На первом уровне архитектуры ИСАОЦСК располагаются распределенные сканеры сетевых ИО. Их задача заключается в сборе ИО,

формировании облака тегов (меток, ярлыков, хештегов, ключевых слов) и приоритезации ИО.

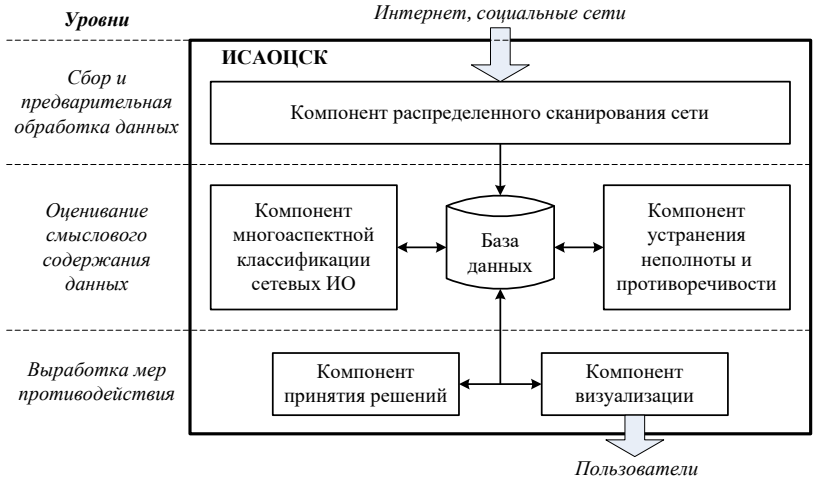


Рис. 1. Общая архитектура ИСАОЦСК

На втором уровне находятся база данных веб-контента, содержащая всю собираемую и обрабатываемую информацию об ИО, а также компонент многоаспектной классификации сетевых ИО и компонент устранения неполноты и противоречивости результатов классификации. На этом уровне исходные данные для классификаторов, сформированные с помощью распределенных сканеров, подвергаются дополнительной обработке с целью устранения неоднозначности (нечеткости) и недостоверности (недостаточности, неполноты).

На третьем уровне располагаются компонент принятий решений, осуществляющий выбор мер противодействия выявленной нежелательной информации, и компонент визуализации результатов работы системы.

Обобщенный алгоритм функционирования ИСАОЦСК можно описать следующим образом.

Шаг 1. Сбор данных о сайтах, потенциально содержащих нежелательную информацию (с помощью компонента распределенного сканирования). Помещение их на хранение в базу данных.

Шаг 2. Выявление и классификация нежелательной информации (компонент многоаспектной классификации). Если классификация про-

шла с высокой точностью, то переход на шаг 4. Иначе — принятие решения о необходимости устранения неполноты и противоречивости собранных данных.

Шаг 3. Функционирование компонента устранения неполноты и противоречивости собранных данных. Переход на шаг 2.

Шаг 4. Выработка и выбор мер противодействия нежелательной информации (компонент принятия решений).

Шаг 5. Визуальное оформление промежуточных и окончательных решений ИСАОЦСК (компонент визуализации).

Рассмотрим решения по построению и функционированию выше указанных компонентов ИСАОЦСК. Компонент визуализации, в силу его специфики, рассматривать не будем.

4. Решения по построению и функционированию компонентов системы. Главная особенность сетевого ИО (СИО), отличающая его от обычного электронного документа, состоит в наличии сложной иерархической структуры. Самым распространенным примером СИО является веб-страница, которая представляет собой набор текстовых файлов, размеченных на языке HTML. Использование HTML позволяет форматировать текст, различать в нём функциональные элементы, создавать гиперссылки и вставлять в отображаемую страницу изображения, звукозаписи и другие мультимедийные элементы. Содержимое веб-страницы называется контентом.

Основная задача компонента распределенного сканирования сети заключается в сборе информации о веб-страницах и предварительной обработке контента. Предлагается использовать в ИСАОЦСК комплекс распределенных интеллектуальных сканеров (КРИС), выполняющих параллельный анализ СИО. Подобный подход подразумевает гибкое масштабирование. Каждый сканер располагается на отдельном хосте и самостоятельно выполняет операции сбора и предобработки сетевого контента.

Исходными данными для распределенного сканирования сети является известное конечное множество X, которое включает в себя n сетевых адресов веб-страниц URL:

$$X = (URL_1, URL_2, \ldots, URL_n). \quad (4)$$

Данное множество распределяется между интеллектуальными сканерами. Каждый сканер загружает результаты сбора и предобработки представленного ему множества объектов в локальное временное хранилище. Далее информация агрегируется в общую базу данных веб-контента.

Архитектура компонента распределенного сканирования сети представлена на рисунке 2.

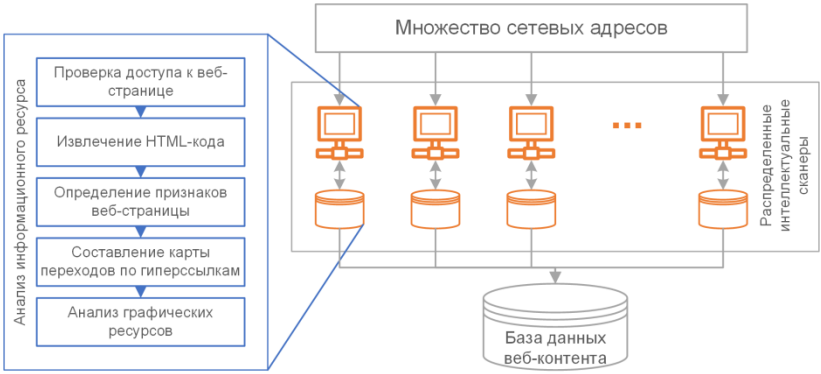


Рис. 2. Компонент распределенного сканирования сети

Для каждого СИО проверяется доступность контента по указанному адресу и присваивается соответствующий статус. При наличии доступа к веб-странице производится загрузка СИО в виде HTML-кода с указанием времени загрузки и его графических ресурсов (изображений, логотипов, элементов фона и т.п.). Формируется множество X', которое содержит HTML-коды для m доступных веб-страниц:

$$X' = (<URL_1, HTML_1>, \ldots, <URL_m, HTML_m>). \quad (5)$$

В случае отсутствия доступа к веб-странице фиксируется код и текст полученной ошибки с указанием времени попытки подключения. Причиной отсутствия доступа может быть устаревание, удаление или некорректный адрес СИО. Также код ошибки может указывать на проблемы с подключением у домена веб-страницы или у сетевого сканера. В этом случае адрес веб-страницы помещается в очередь на повторную проверку.

HTML-код веб-страницы используется для извлечения признаков виде конечного множества X'' размерностью m:

$$X'' = (<URL_1, a_{11},\ldots,a_{k1}>, \ldots, <URL_m, a_{1m},\ldots,a_{km}>), \quad (6)$$

где a_{ij} — i-й признак ($i = 1, \ldots, k$) j-ой веб-страницы ($j = 1, \ldots, m$).

В качестве признаков веб-страниц предлагается использовать:

– тип контента (текст, изображение, видео, аудио и т.д.);

– размер контента;

– текстовое содержимое;

– язык текстового содержимого;

– длину текстового содержимого;

– количество гиперссылок на внутренние СИО (принадлежащие к тому же домену);

– количество гиперссылок на внешние СИО;

– количество графических ресурсов.

Текстовое содержимое СИО представляется в виде описания блоков текста с их частотной характеристикой. Информация о взаимосвязи нескольких СИО отображается в виде карты переходов. Для каждого i-го корневого СИО определяется следующее множество:

$$H_i = (<h_{i1}, d_{i1}, l_{i1}>, \ldots, <h_{is}, d_{is}, l_{is}>), \quad (7)$$

где h_{ij} ($j = 1, \ldots, s$) — сетевой адрес j-го стороннего СИО, извлекаемого из гиперссылок на веб-странице i-го корневого СИО; d_{ij} — глубина взаимосвязи двух СИО, обозначающая число переходов от корневого СИО до текущего; l_{ij} — показатель локальности, определяющий, связан корневой СИО с внутренним ресурсом (принадлежащим тому же домену) или внешним (принадлежащим стороннему домену).

В процессе сбора данных о СИО происходит выгрузка и последующий анализ изображений, содержащихся в HTML-коде и в таблице стилей. Стили хранятся в отдельном CSS-файле, который может быть использован для любых информационных ресурсов одного домена. Предлагается выделять следующие признаки графических ресурсов:

– сетевой адрес изображения;

– сетевой адрес домена, на котором хранится изображение;

– наименование;

– графический формат файла (JPEG, JPG, PNG, GIF и т.д.);

– цветовую модель (RGB, RGBA, CMYK);

– ширину изображения в пикселях;

– длину изображения в пикселях;

– список основных цветов изображения (центроидов цветовых кластеров) и соответствующий им процент принадлежащих пикселей в формате {«код цвета» : «процент пикселей»};

– текст на изображении (если он присутствует) в формате {«код языка» : «текст»}.

Следует отметить, что на текущем этапе разработки ИСАОЦСК в состав признаков текстовых и графических ресурсов пока еще не входят признаки, характеризующие интерпретацию ИО. В результате возможны ложноположительные ошибки, при которых к категории нежелательной информации могут быть отнесены, например, сайты с наборами данных, содержащими примеры сетевых атак. Охват признаков интерпретации ИО рассматривается как направление дальнейших исследований.

Таким образом, компонент распределенного сканирования сети реализует следующие функции: (1) обнаружение и загрузка веб-контента для заданного множества сетевых адресов; (2) структурная категоризация СИО; (3) вычисление частотных характеристик для структурных элементов веб-страниц; (4) построение карт переходов СИО.

Компонент многоаспектной классификации сетевых ИО включает в свой состав четыре модуля:

1) модуль фильтрации содержимого СИО;
2) модуль извлечения признаков из СИО;
3) модуль предобработки признаков СИО и построения обучающих выборок;
4) модуль классификации СИО.

Предназначение модуля фильтрации содержимого СИО заключается в удалении знаков препинания, а также тех единиц речи, представленных в виде отдельных слов и словосочетаний, которые не влияют на смысловое наполнение текста. К таким словам относятся местоимения, предлоги, цифры, артикли, модальные глаголы и союзы, а также те слова, которые являются свойственными одновременно для нескольких классов и используемыми в различных контекстах (например, «теперь», «тогда», «только»). Кроме этого, из рассмотрения исключаются данные, помещенные в секции комментариев, поскольку они не видны конечному пользователю.

Модуль извлечения признаков из СИО построен на основе DOM (Document Object Model) парсинга, что позволяет быстро и удобно извлекать из исходной html-страницы текст, заключенный в искомый тег. В данном модуле реализована поддержка вычисления параметров html-страницы с использованием трех типов исходных данных: структуры документа, текстового содержимого, а также URL-строки. В таблице 1 перечислены наименования html-тегов, частоты встречаемости которых формируют параметры вектора признаков в соответствии с типом исходных данных «Структура документа».

При формировании параметров, соответствующих типу исходных данных «Текстовое содержимое», применялся подход на основе

«мешка слов». Исходный документ $T = \{w_1, \ldots, w_N\}$ представляется последовательностью слов. Он преобразуется в список пар $L=\{(w_1, Q(w_1, T)), \ldots, (w_M, Q(w_M, T))\} = \{(w_1, q_1), \ldots, (w_M, q_M)\}$, где элементы $\{w_1, \ldots, w_M\} = D \subset T$ являются уникальными словами ($M \leq N$) (множество D называется словарем), а элементы $\{q_1 = Q(w_1, T), \ldots, q_M = Q(w_M, T)\}$ отражают абсолютные частоты появления соответствующих слов в документе T (Q — функция, возвращающая число вхождений слова, представленного первым аргументом, в документ, представленный вторым аргументом).

Таблица 1. Описание html-тегов, извлекаемых из сетевого ИО

№	html-тег	Описание html-тега
1	<b>	Жирное выделение текста
2	<dt>	Создание элемента в списке определений
3		Разбиение документа на фрагменты
4	<h1>	Задание заголовка первого уровня
5	<h2>	Задание заголовка второго уровня
6	<h3>	Задание заголовка третьего уровня
7	<h4>	Задание заголовка четвертого уровня
8	<h5>	Задание заголовка пятого уровня
9	<h6>	Задание заголовка шестого уровня
10	<link>	Задание связи с внешним ресурсом
11	<a>	Создание ссылки
12	<form>	Задание формы
13		Создание элемента маркированного или нумерованного списка
14	<i>	Курсивное выделение текста
15	<p>	Выделение абзаца

Оценка семантической схожести двух документов сводится к вычислению количества слов, которые одновременно встречаются в этих документах. Чем меньшее количество общих слов содержится в обоих документах, тем ниже их уровень семантической схожести. Следует отметить, что с увеличением объема документа модель «мешка слов» становится особенно чувствительной к сравнению документов. В этом случае для уменьшения временных затрат рассматривается множество $L' = \{(w, q) \mid (w, q) \in L \wedge q > h \geq 0\}$ вместо L, что позволяет игнорировать редко встречающиеся слова, частота вхождения которых в документ не превосходит заранее заданного числового порога h.

Среди ограничений, присущих данной модели, следует отметить невозможность учета контекста, в котором может использоваться то или иное слово. Для устранения этого недостатка прибегают к построе-

нию матрицы семантических связей [34]. С этой целью разработан алгоритм вычисления семантической схожести наиболее употребительных слов на основе модели word2vec [35] (рис. 3).

В данном алгоритме для каждого класса S_i формируется документ R, объединяющий все документы этого класса. Для документа R вычисляется функция *find_common_words* (R, d), которая извлекает d слов, наиболее часто встречающихся внутри R. Результат применения данной функции обозначается как W. После этого внутри каждого документа класса S_i находятся c наиболее употребительных слов, для которых вычисляется семантическая схожесть с каждым словом из набора W. С этой целью применяется функция *word2vec*. Полученные результаты записываются в виде компонентов вектора признаков, соответствующих типу исходных данных «Текстовое содержимое».

Входные данные: $\Omega = \{S_1, S_2, S_3, ...\} = \{arts, business, computers, ...\}$ — классы документов
$\Phi = \left\{\{T_{ij}\}_{j=1}^{N_i}\right\}_{i=1}^{\#\Omega}$ — набор документов с разделением по классам
$c = 3$ — количество наиболее употребительных слов внутри документа
$d = 5$ — количество наиболее употребительных слов внутри совокупности документов, принадлежащих одному и тому же классу

Выходные данные: $\Psi = \left\{\{F_{ij}\}_{j=1}^{N_i}\right\}_{i=1}^{\#\Omega}$ — набор признаков документов

для каждого $i \in \{1, \ldots, \#\Omega\}$ **выполнять**
  $R \coloneqq \varnothing$
  для каждого $j \in \{1, \ldots, N_i\}$ **выполнять**
    $R \coloneqq R \cup T_{ij}$
  $W \coloneqq find\_common\_words(R, d)$
  для каждого $j \in \{1, \ldots, N_i\}$ **выполнять**
    $F_{ij} \coloneqq \varnothing$
    для каждого $v \in find\_common\_words(T_{ij}, c)$ **выполнять**
      для каждого $w \in W$ **выполнять**
        $F_{ij} \coloneqq F_{ij} \cup \{word2vec(v, w)\}$

Рис. 3. Алгоритм вычисления семантической схожести наиболее употребительных слов

Если в качестве исходных данных используются URL-строки, то тогда для формирования признаков СИО проверяется признак вхождения в эти строки десяти наиболее употребительных слов, характерных для каждой категории.

В модуле предобработки признаков СИО и построения обучающих выборок реализована поддержка минимаксной нормализации и разбиения исходной выборки на тестовую и обучающую части.

В функционировании модуля классификации СИО выделяются два режима: режим обучения и режим анализа. В первом режиме выполняется настройка классификаторов путем итеративного предъявления на их вход последовательностей обучающих векторов и последующей корректировки внутренних параметров классификаторов. Во втором режиме осуществляется выделение класса анализируемого ИО, включая характер и степень вредоносности сетевого контента.

Компонент устранения неполноты и противоречивости результатов классификации СИО предназначен для устранения неопределенности (нечеткости, неполноты и противоречивости) оценки СИО. Такая оценка почти во всех случаях осуществляется в условиях неопределенности исходных данных — измеряемых, моделируемых или наблюдаемых в шумах атрибутов СИО (текстовых, графических, числовых, булевых, ординальных, номинальных и т.д.). Полагается, что основным источником такой неопределенности является «Текстовое содержимое» веб-страниц. При этом доминирующими видами неопределенности являются неоднозначность (нечеткость, противоречивость) и недостаточность (неполнота) исходных данных.

Неопределенность оценки СИО вызвана нестационарностью поступления информации, нечеткостью, неполнотой и противоречивостью идентификации признаков такой информации, динамикой функционирования системы защиты и воздействиями дестабилизирующих (зачастую антагонистических) факторов внешней среды, а также неопределенностью целей и несогласованностью задач обнаружения и противодействия нежелательной информации.

Общий алгоритм функционирования компонента устранения неполноты и противоречивости результатов классификации СИО опирается на модели и механизмы устранения неопределенности оценки признаков нежелательной информации (в интересах ее обнаружения и противодействия ей). Он использует методы обработки нечетких, неполных и противоречивых знаний и включает два ключевых этапа.

В основе *первого этапа*, ориентированного на устранение неопределенности классификации СИО на основе нечетких множеств [31, 32], лежит механизм поддержки принятия решения по включению (либо не включению) нечетко заданных признаков информации, циркулирующей в цифровом сетевом контенте, во множество признаков нежелательной информации. Иными словами, если в СИО наличие,

объем и номенклатура (уровень опасности) признаков какой либо сомнительной информации превышает допустимый порог (α-уровень функции принадлежности), то эта информация оценивается и классифицируется как нежелательная, потенциально вредоносная. При этом субъективная мера уверенности, с которой данная информация принадлежит нечеткому множеству признаков нежелательной информации, задается функциями принадлежности. При этом для объединения нескольких субъективных мер уверенности (мнений нескольких экспертов) используются математические операции дополнения, объединения, пересечения нечетких множеств и операция дизъюнктивного суммирования нечетких множеств.

Потребность устранять нечеткость признаков анализируемой информации возникает на фоне того, что данные признаки фактически определены, сформулированы, но их значения заданы нечетко. Эти значения поступают из множества разнородных источников и могут неоднозначно, с помощью нечетких высказываний (лингвистических термов типа «много», «сильно» и т.п.), указывать, например, на меру уверенности, с которой конкретный контент анализируемой веб-страницы принадлежит (либо не принадлежит) множеству признаков нежелательной информации.

В основе *второго этапа*, ориентированного на устранение неопределенности оценки и категоризации на основе искусственной нейронной сети (ИНС), лежит нейросетевая модель [28, 30] поиска и прогнозирования неполно и противоречиво заданных признаков анализируемой информации. Эта модель позволяет осуществлять поиск взаимосвязей между признаками и обоснованно включать (либо не включать) неполно и противоречиво заданные признаки во множество признаков нежелательной информации. Иными словами, если есть хотя бы один признак, гарантированно включаемый в состав множества признаков нежелательной информации, то можно построить такой вектор входных признаков, который учитывает неполные и противоречивые взаимосвязи всех признаков (по мнению экспертов). Тогда с помощью ИНС можно получить выходной вектор признаков с коэффициентами, характеризующими их вес (уровень опасности) и, в свою очередь, оценить и классифицировать эту информацию как нежелательную.

Итоговым результатом работы компонента устранения неполноты и противоречивости является окончательно сформулированная система признаков СИО, однозначно определяющая принадлежность (либо не принадлежность) конкретной информации к нежелательной, с учетом устранения неопределенности в рамках моделей и алгоритмов обработки нечетких, неполных и противоречивых знаний. При этом

предложенный подход позволяет работать с обоими видами неопределенности раздельно, что сокращает объем производимых вычислений и приводит к увеличению быстродействия ИСАОЦСК для защиты от нежелательной информации.

Компонент принятия решений по противодействию нежелательной информации использует в своей работе теорию принятия решений, включая методы многокритериальной оптимизации. На вход компонента поступают: (1) нежелательные СИО; (2) доступные контрмеры. Основными этапами работы компонента являются: (1) создание моделей СИО, информационной системы, противодействия и процесса противодействия; (2) выбор контрмер. На выходе компонента формируется набор выбранных мер противодействия.

Информационной системой, в которой реализовано противодействие, является Интернет. Модель информационной системы задается следующим образом: $IS = (IO, IC)$, где *IO* — сетевые ИО, *IC* — связывающие их коммуникационные средства.

Модель СИО определяется следующим образом:

$$IO = <size, role, hltype, type, state, ioaud, saud>, \quad (8)$$

где *size* — размер СИО, может иметь значения из множества {*sm*, *mi*, *la*}, *sm* — «малый», *mi* — «средний», *la* — «большой»;

role — роль СИО, может иметь значения из множества {*s*, *r*, *u*}, *s* — «отправитель», *r* — «получатель», *u* — «пользователь»;

hltype — абстрактный тип СИО, принимает значение *h*, если СИО является нежелательным, и *n* — в противном случае;

type — детальный тип СИО, может принимать значения из множества {*ter*, *hea*, *por*, *dru*, *cru*, *none*}, *ter* — СИО, содержащий призывы к терроризму и экстремизму; *hea* — СИО, содержащий информацию, вредную для здоровья людей (особенно детей), морального и духовного развития; *por* — СИО с пропагандой порнографии; *dru* — СИО, содержащий информацию о путях распространения наркотиков и призывы к суициду; *cru* — СИО, содержащий призывы к насилию (войне); *none* — СИО не является нежелательным (*hltype* равен *n*);

state — состояние компрометации СИО, может принимать значения *compr*, если СИО скомпрометирован вредоносной информацией, и *nonc* — если не скомпрометирован;

ioaud — аудитория СИО, представляющая собой массив ссылок, которые связаны с отправителем посредством сообщений и которые являются получателями объектов (может быть нулевым);

saud — вещественное число (при наличии счетчика посетителей СИО) или экспертная оценка субъектов, являющихся получателями СИО (может быть 0).

Модель контрмеры *rm* из множества контрмер *RM* задается в следующем виде:

$$rm = <rm\_class, rm\_type, rm\_cost, rm\_role, rm\_ef, rm\_cd>, \quad (9)$$

где *rm_class* — класс контрмеры (барьер, маскировка, информирование или принуждение); *rm_type* — размер СИО (малый, средний или большой); *rm_cost* — стоимость контрмеры; *rm_role* — роль СИО; *rm_ef* — эффективность контрмеры; *rm_cd* — побочный ущерб от реализации контрмер.

Модель контрмеры используется для определения модели противодействия. Противодействие влияет на состояние информационной системы: {*IO*, *IC*} становится $\{IO^l, IC^l\}$, где *l* — номер контрмеры. Для *j* информационных объектов из IO^l ($j = 0,\ldots,N$, где *N* – номер элемента в *ioaud* нежелательного СИО, на которого воздействует контрмера), СИО удаляется, или модифицируются следующие их параметры: *role* принимает значение *r* или *u*, *hltype* становится равным *n*, *type* становится равным *none*, *state* становится равным *nonc*. Для *d* связей из IC^l ($d = 0,\ldots,D$, где *D* — номер связи между нежелательным СИО и связанным объектом, на который воздействует контрмера) информационная связь удаляется.

Модели (8) и (9) используются для формализации алгоритма противодействия нежелательной информации. Входными данными этого алгоритма являются: размер СИО, роль СИО, абстрактный тип СИО (параметр *hltype*) и детальный тип СИО (параметр *type*). Алгоритм включает две фазы. На первой фазе производится анализ аудитории нежелательной информации. На второй производится анализ и выбор меры противодействия. Для учета аудитории нежелательной информации на первой фазе производится поиск связанных объектов и изменение их состояния на скомпрометированное. Затем, с учетом этих скомпрометированных объектов и их трафика (с помощью счетчиков), вычисляются размер и возраст аудитории.

Анализ контрмеры на второй фазе заключается в вычислении ее эффективности (параметр *rm_ef*) и стоимости (*rm_cost*). Эффективность вычисляется как отношение СИО-получателей, которые не будут скомпрометированы в случае реализации противодействия, к общему количеству получателей. Следует отметить, что при оценке эффективности не учитываются возможные случаи самокомпрометирования, когда получатель попадает под действия средств защиты, например, ловушек;

учет таких случаев относится к дальнейшим исследованиям. Стоимость задается экспертами вручную. Кроме того, учитываются класс средств противодействия (который выбирается в зависимости от типа вредоносной информации) и размер информационного объекта.

Для выбора контрмеры на второй фазе используются предварительно сформированные правила. В качестве примера приведем одно из правил выбора контрмеры, используемое в разработанном алгоритме и основанное на параметрах моделей (8) и (9): «если *role = s* и *size = sm* и *type = ter* и размер аудитории меньше 3000 и возраст аудитории больше 18, то тогда выбрать контрмеры, у которых *rm_class* равно *disguise* или *informing* либо *rm_type* равно *small*».

5. Реализация и экспериментальная оценка системы. Для проведения экспериментальной оценки системы с помощью компонента распределенного сетевого сканирования был сформирован набор данных, содержащий категорированный веб-контент. Интеллектуальные сканеры были размещены на четырех хостах. Характеристика вычислительной базы сканеров приведена в таблице 2.

Таблица 2. Характеристика вычислительной базы сканеров

№	Процессор	Тактовая частота (ГГц)	ОЗУ (Гб)	ОС
1	Intel Core i5-8250U	1,8	8	Windows 10
2	Intel Core i7-7700HQ	2,8	12	Windows 10
3	Inter Core i7-8665U	1,9	16	Windows 10
4	AMD Ryzen 5 3500U	2,1	8	Windows 10

Исходный набор данных, представляющий множество сетевых адресов, получен из общедоступных категорированных списков веб-страниц, включающих в себя URLBlacklist, MESD blacklists [36], Shallalist [37] и DMOZ [38]. В нем содержатся адреса веб-страниц, маркированных 23 категориями контента, в том числе относящимися к нежелательной информации (таблица 3).

Для создания сбалансированного набора множества сетевых адресов была проверена доступность веб-ресурсов различных категорий, так как открытые категорированные списки могут содержать много устаревших данных. Для проведения эксперимента было введено ограничение на 2000 доступных веб-страниц для каждой категории, за исключением «алкоголь» и «политика», для которых в исходных списках содержится меньшее количество маркированных данных. Итоговое экспериментальное множество адресов веб-страниц включает в себя 44 866 URL, информация о которых записана в общую базу данных. Для каждой веб-страницы определены следующие признаки:

– *id* – идентификатор веб-страницы;

– *url* – сетевой адрес веб-страницы;

– *category* – категория веб-страницы;

– *domain* – домен веб-страницы;

– *status* – доступность веб-страницы;

– *content_type* – тип контента;

– *content_length* – размер контента;

– *language* – язык текстового содержимого веб-страницы;

– *text_length* – размер текстового содержимого веб-страницы;

– *local_hyperlinks_count* – количество гиперссылок на ресурсы того же домена, содержащиеся на веб-странице.

– *external_hyperlinks_count* – количество гиперссылок на внешние ресурсы, содержащиеся на веб-странице.

Таблица 3. Категории веб-контента

No.	Название категории	Нежелательная информация	Число веб-страниц
1	Для взрослых (adult)	✓	2000
2	Агрессия (aggression)	✓	2000
3	Алкоголь (alcohol)	✓	1386
4	Искусство (arts)	×	2000
5	Бизнес (business)	×	2000
6	Компьютеры (computers)	×	2000
7	Сервисы знакомств (dating)	✓	2000
8	Наркотики (drugs)	✓	2000
9	Игры (games)	×	2000
10	Азартные игры (gamling)	✓	2000
11	Хакерство (hacking)	✓	2000
12	Медицина (health)	×	2000
13	Дом (home)	×	2000
14	Для детей (kids)	×	2000
15	Новости (news)	×	2000
16	Политика (politics)	×	1480
17	Досуг (recreation)	×	2000
18	Ссылки (reference)	×	2000
19	Религия (religion)	✓	2000
20	Наука (science)	×	2000
21	Шоппинг (shopping)	×	2000
22	Общество (society)	×	2000
23	Спорт (sports)	×	2000

Пример отображения перечисленных признаков в базе данных представлен на рисунке 4.

id	url	category	domain	status	content_type	content_length	language	text_length	local_hyperlinks_count	external_hyperlinks_count
2771887	http://feeds.fee…	Arts	feedb…	OK	text/xml; char…	21732	en	21692	1	0
2771893	http://www.npr.…	Arts	npr.org	OK	text/xml;chars…	4919	en	19181	1	0
2771895	http://www.artis…	Arts	artistdi…	OK	text/html; char…	73785	en	73735	89	11
2771896	http://www.curli…	Arts	curlio.…	OK	text/html	49610	en	49610	176	1
2771897	http://www.cbc.…	Arts	cbc.ca	OK	text/html; char…	13570	en	44968	41	15
2771898	http://www.musi…	Arts	musicn…	OK	text/html; char…	88598	en	88598	1	63
2771899	http://www.xfm.…	Arts	xfm.co…	OK	text/html; char…	166234	en	166216	257	29
2771900	http://www.mi2…	Arts	mi2n.com	OK	text/html; char…	159297	en	159136	34	22
2771901	http://www.ukm…	Arts	ukmusi…	OK	text/html; char…	31368	en	31367	1	42
2771903	http://www.xs4…	Arts	xs4all.nl	OK	text/html	9836	en	24186	8	1
2771904	http://www.anti…	Arts	antimu…	OK	text/html	71920	en	71920	54	207
2771906	http://nextbigthi…	Arts	nextbi…	OK	text/html; char…	1389359	en	1389358	3981	104

Рис. 4. Признаки сетевых информационных объектов

Общий размер извлеченных HTML-кодов веб-страниц составляет 3,3 Гбайта. На рисунке 5 представлен размер текстового контента для каждой категории: количество строк таблицы в базе данных (*Rows*), средняя длина строки (*Avg Row Length*) и размер данных (*Data Length*). Объем всей базы данных составляет 8,26 Гбайта.

Name	Engine	Version	Row Format	Rows	Avg Row Length	Data Length
text_adult	InnoDB	10	Dynamic	546465	196	102.6 MiB
text_aggression	InnoDB	10	Dynamic	632059	291	175.6 MiB
text_alcohol	InnoDB	10	Dynamic	508918	320	155.6 MiB
text_arts	InnoDB	10	Dynamic	381114	356	129.6 MiB
text_business	InnoDB	10	Dynamic	385600	289	106.6 MiB
text_computers	InnoDB	10	Dynamic	507244	478	231.7 MiB
text_dating	InnoDB	10	Dynamic	606009	271	156.6 MiB
text_drugs	InnoDB	10	Dynamic	531781	257	130.6 MiB
text_gambling	InnoDB	10	Dynamic	228279	576	125.6 MiB
text_games	InnoDB	10	Dynamic	252228	405	97.6 MiB
text_hacking	InnoDB	10	Dynamic	225028	347	74.6 MiB
text_health	InnoDB	10	Dynamic	573686	289	158.6 MiB
text_home	InnoDB	10	Dynamic	591596	311	175.6 MiB
text_kids	InnoDB	10	Dynamic	380076	1999	724.6 MiB
text_news	InnoDB	10	Dynamic	548568	381	199.7 MiB
text_politics	InnoDB	10	Dynamic	369384	481	169.6 MiB
text_recreation	InnoDB	10	Dynamic	387911	371	137.6 MiB
text_reference	InnoDB	10	Dynamic	576658	332	182.6 MiB
text_religion	InnoDB	10	Dynamic	513433	303	148.6 MiB
text_science	InnoDB	10	Dynamic	361502	1234	425.6 MiB
text_shopping	InnoDB	10	Dynamic	462285	244	107.6 MiB
text_society	InnoDB	10	Dynamic	494495	330	155.6 MiB
text_sports	InnoDB	10	Dynamic	440567	306	128.6 MiB

Рис. 5. Размер текстового содержимого веб-страниц

Следует отметить, что анализ графических ресурсов веб-страниц выходил за рамки проведенных исследований. Основное внимание уделялось признакам СИО и их текстовому содержимому.

Компонент многоаспектной классификации СИО оценивался по четырем показателям: достоверности (*accuracy*), точности (*precision*), полноте (*recall*) и F-мере (*F-measure*). Использовалось семь классификаторов: простой мешок слов; взвешенный мешок слов; непрерывный

мешок слов, классификатор skip-gram, сверточная нейронная сеть, классификатор fastText [39] и случайный классификатор.

В таблице 4 представлены указанные показатели, вычисленные для каждого классификатора.

Таблица 4. Показатели достоверности, точности, полноты и F-меры, вычисленные для трехблочной кросс-валидации

Классификатор / Показатель	Простой мешок слов	Взвешенный мешок слов	Непрерывный мешок слов	Классификатор skip-gram	Сверточная нейронная сеть	Классификатор fastText	Случайный классификатор
Достоверность	63,16%	64,69%	16,02%	22,53%	25,81%	84,15%	5,28%
Точность	63,06%	66,2%	26,66%	36,8%	29,56%	80,48%	5,28%
Полнота	58,63%	61,04%	12,21%	18,53%	20,42%	78,83%	5,28%
F-мера	60,76%	63,52%	16,75%	24,65%	24,15%	79,65%	5,28%

Показатели точности, полноты и F-меры вычислялись для каждого класса в отдельности. Поэтому значения соответствующих показателей были усреднены по всем классам. Случайный классификатор с равной вероятностью генерировал числовую метку одного из 19 классов нежелательной информации. Эксперимент для этого классификатора проводился 100 раз, что позволило достаточно точно приблизить экспериментальные (5,28%) и теоретические (1/19 * 100 % ≈ 5,26 %) оценки показателей. Наилучшие результаты показал классификатор fastText (его достоверность равна 84,15%). На рисунке 6 представлены детальные показатели точности, полноты и F-меры, вычисленные для этого классификатора с использованием трехблочной кросс-валидации.

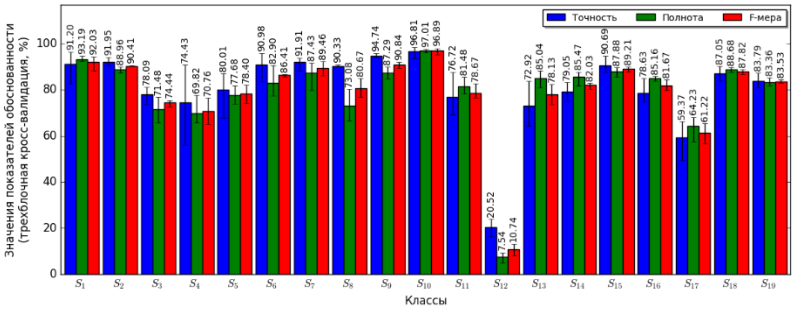


Рис. 6. Значения показателей точности, полноты и F-меры, вычисленные для классификатора fastText и трехблочной кросс-валидации

На рисунке 7 показана зависимость количества ошибок, показателя достоверности и среднеквадратичной ошибки сверточной нейронной сети от номера эпохи обучения. В экспериментах использовалась нейронная сеть с двумя слоями свертки (с функцией активации ReLU) и следующим за каждым из них слоем субдискретизации (с функцией max). Достоверность на обучающей выборке для нейронной сети не превосходит 32,55%. Этим в полной мере объясняется низкое значение соответствующего показателя (25,81%) на тестовой выборке.

Таким образом, эксперименты показали, что максимальная эффективность обнаружения нежелательной информации, определяемая показателями достоверности, точности, полноты и F-меры, достигается в ИСАОЦСК при использовании классификатора fastText. Однако исследования в области повышения эффективности будут продолжаться и дальше.

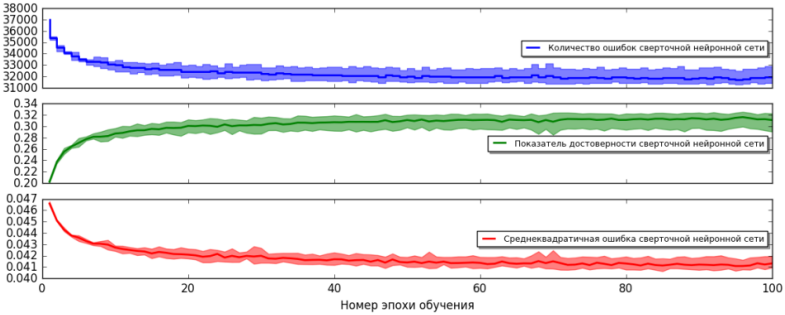


Рис. 7. Зависимость количества ошибок, показателя достоверности и среднеквадратичной ошибки сверточной нейронной сети от номера эпохи обучения

Здесь представляет интерес подход, основанный на комплексировании различных классификаторов, предложенный в [40]. При таком подходе итоговая эффективность классификации становится выше, чем эффективность отдельного классификатора, подлежащего комплексированию. Для того чтобы реализовать этот подход в ИСАОЦСК, необходимо проверить работу различных схем комплексирования (мажоритарной, взвешенной и т.д.). Авторы относят это к направлениям дальнейших исследований.

Рассмотрим теперь примеры реализации компонента устранения неполноты и противоречивости и связанного с ним компонента принятия решений. Для реализации первого этапа, ориентированного на использование нечетких множеств, используются методы обработки не-

четких знаний (вычисления дизъюнктивной суммы). Положим, что задан исходный состав множества нечетко заданных признаков и сформулированы нечетко заданные мнения экспертов — начальные функции принадлежности нечетких множеств, характеризующие предварительный, нечетко заданный состав множества признаков СИО:

$$\tilde{\tilde{J}} = [\Delta\tilde{\tilde{j}}_{\text{терр}} \big| \mu(\Delta\tilde{j}_{\text{терр}});\ \Delta\tilde{\tilde{j}}_{\text{дет}} \big| \mu(\Delta\tilde{j}_{\text{дет}});\ \Delta\tilde{\tilde{j}}_{\text{порн}} \big| \mu(\Delta\tilde{j}_{\text{порн}});\ \Delta\tilde{\tilde{j}}_{\text{нарк}} \big| \mu(\Delta\tilde{j}_{\text{нарк}});\ \Delta\tilde{\tilde{j}}_{\text{войн}} \big| \mu(\Delta\tilde{j}_{\text{войн}})]^{\text{т}}, \quad (10)$$

где $\Delta\tilde{j}_{\text{терр}}(k)$ — признак СИО, характеризующий аномальное отклонение в трафике среднего количества информации, содержащей публичные призывы к осуществлению террористической и экстремистской деятельности; $\Delta\tilde{j}_{\text{дет}}(k)$ — признак СИО, характеризующий аномальное отклонение в контенте среднего количества информации, причиняющей вред здоровью, нравственному и духовному развитию людей (особенно детей); $\Delta\tilde{j}_{\text{порн}}(k)$ — признак СИО, характеризующий аномальное отклонение среднего количества информации, нацеленной на пропаганду порнографии; $\Delta j_{\text{нарк}}(k)$ — признак СИО, характеризующий аномальное отклонение среднего количества информации, содержащей данные о способах разработки, изготовления и использования наркотических средств и совершения самоубийства, а также нецензурную брань, а $\Delta\tilde{j}_{\text{войн}}(k)$ — признак СИО, характеризующий аномальное отклонение в контенте среднего количества прямых призывов к насилию и жестокости (войне), этнической и религиозной ненависти либо вражде; символ μ — функция принадлежности нечеткого множества, принимающая значения от 0 до 1.

Дизъюнктивная сумма двух нечетких множеств $\tilde{X}$ и $\tilde{Y}$, характеризующих мнения первого и второго экспертов о степени принадлежности признаков СИО к множеству опасных признаков, имеет следующий вид:

$$\tilde{X} \oplus \tilde{Y} = (\tilde{X} \cap \bar{\tilde{Y}}) \cup (\bar{\tilde{X}} \cap \tilde{Y}), \quad (11)$$

где $\bar{\tilde{X}}$ и $\bar{\tilde{Y}}$ — дополнения этих нечетких множеств.

Тогда функция принадлежности для *i*-го признака имеет вид:

$$\forall j_i \in \overline{1,\ldots,5}: \quad \begin{array}{l} \mu_{\tilde{X}\oplus\tilde{\bar{Y}}}(j_i) = \max\{[\min\{\mu_{\tilde{X}}(j_i), 1-\mu_{\tilde{Y}}(j_i)\}]; \\ \quad [\min\{1-\mu_{\tilde{X}}(j_i), \mu_{\tilde{Y}}(j_i)\}]\}. \end{array}, \qquad (12)$$

Мнение первого (X) из двух экспертов об оценке и категоризации каждого признака как признаков нежелательной информации можно представить в виде следующего нечеткого множества:

$$\tilde{X} = \{\Delta\tilde{j}_{\text{терр}}|\,0{,}3;\ \Delta\tilde{j}_{\text{дет}}|\,0{,}1;\ \Delta\tilde{j}_{\text{порн}}|\,0{,}1;\ \Delta\tilde{j}_{\text{нарк}}|\,0{,}5;\ \Delta\tilde{j}_{\text{войн}}|\,0{,}2\}.$$

Аналогичное мнение второго (Y) эксперта можно представить в виде следующего нечеткого множества:

$$\tilde{Y} = \{\Delta\tilde{j}_{\text{терр}}|\,0{,}7;\ \Delta\tilde{j}_{\text{дет}}|\,0{,}9;\ \Delta\tilde{j}_{\text{порн}}|\,0{,}4;\ \Delta\tilde{j}_{\text{нарк}}|\,0{,}5;\ \Delta\tilde{j}_{\text{войн}}|\,0{,}4\}.$$

Для этих нечетких множеств их дополнения равны:

$$\bar{\tilde{X}} = \{\Delta\tilde{j}_{\text{терр}}|\,0{,}7;\ \Delta\tilde{j}_{\text{дет}}|\,0{,}9;\ \Delta\tilde{j}_{\text{порн}}|\,0{,}9;\ \Delta\tilde{j}_{\text{нарк}}|\,0{,}5;\ \Delta\tilde{j}_{\text{войн}}|\,0{,}8\};$$

$$\bar{\tilde{Y}} = \{\Delta\tilde{j}_{\text{терр}}|\,0{,}3;\ \Delta\tilde{j}_{\text{дет}}|\,0{,}1;\ \Delta\tilde{j}_{\text{порн}}|\,0{,}6;\ \Delta\tilde{j}_{\text{нарк}}|\,0{,}5;\ \Delta\tilde{j}_{\text{войн}}|\,0{,}6\},$$

а пересечения этих нечетких множеств имеют вид:

$$\tilde{X} \cap \bar{\tilde{Y}} = \{\Delta\tilde{j}_{\text{терр}}|\,0{,}3;\ \Delta\tilde{j}_{\text{дет}}|\,0{,}1;\ \Delta\tilde{j}_{\text{порн}}|\,0{,}1;\ \Delta\tilde{j}_{\text{нарк}}|\,0{,}5;\ \Delta\tilde{j}_{\text{войн}}|\,0{,}2\};$$

$$\bar{\tilde{X}} \cap \tilde{Y} = \{\Delta\tilde{j}_{\text{терр}}|\,0{,}7;\ \Delta\tilde{j}_{\text{дет}}|\,0{,}9;\ \Delta\tilde{j}_{\text{порн}}|\,0{,}4;\ \Delta\tilde{j}_{\text{нарк}}|\,0{,}5;\ \Delta\tilde{j}_{\text{войн}}|\,0{,}4\}.$$

В итоге объединение этих нечетких множеств дает следующие конечные результаты дизъюнктивного суммирования, характеризующие совокупное мнение двух экспертов об оценке и категоризации каждого признака как признака нежелательной информации:

$$\tilde{X} \oplus \tilde{Y} = (\tilde{X} \cap \bar{\tilde{Y}}) \cup (\bar{\tilde{X}} \cap \tilde{Y}) =$$
$$= \{\Delta\tilde{j}_{\text{терр}}|\,0{,}7;\ \Delta\tilde{j}_{\text{дет}}|\,0{,}9;\ \Delta\tilde{j}_{\text{порн}}|\,0{,}4;\ \Delta\tilde{j}_{\text{нарк}}|\,0{,}5;\ \Delta\tilde{j}_{\text{войн}}|\,0{,}4\}.$$

В случае, когда экспертов больше двух, формулируется мнение третьего эксперта, итоговое совокупное мнение двух первых экспертов выступает в роли отдельного мнения, и цикл повторяется заново до тех пор, пока не иссякнут эксперты. Тогда получим совокупное, единое мнение экспертов на основе обработки нечетких знаний.

Критерием оценки СИО в рассмотренном случае выступает «граничное», пороговое значение функции принадлежности, описывающей важность (предпочтительность) включения признаков СИО в состав множества признаков нежелательной информации, например, на уровне $\mu^{\text{тр}} \geq 0{,}65$.

Завершающим шагом экспериментальной оценки признаков СИО в условиях неопределенности для данного этапа является отбор конкретных признаков в состав множества признаков нежелательной информации. Графики значений функции принадлежности, описывающей критерий оценки и категоризации в условиях нечеткости, полученные для рассмотренного примера, представлены на рисунке 8.

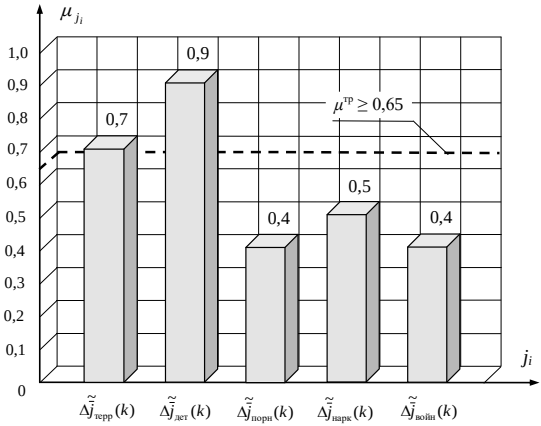


Рис. 8. Результаты вычислительного эксперимента по оценке и категоризации признаков нежелательной информации в условиях нечеткости

Из рисунка 8 видно, что признаки $\Delta\bar{j}_{\text{порн}}(k)$ $\Delta j_{\text{нарк}}(k)$ и $\Delta\bar{j}_{\text{войн}}(k)$ не превосходят опасный уровень и не являются нежелательными и потенциально вредоносными. Предпочтение по опасности отдано $\Delta\bar{j}_{\text{терр}}(k)$ и $\Delta\bar{j}_{\text{дет}}(k)$, то есть аномальному отклонению среднего количества информации, содержащей призывы к терроризму и причиняющей вред здоровью детей. Расчеты характеризуют вес (важность, предпочтительность) конкретного нечеткого признака. Предложенные значения функций принадлежности можно интерпретировать как прогноз гарантированной предпочтительности включения конкретного признака в состав множества признаков нежелательной информации.

Рассмотрим второй пример, демонстрирующий возможности компонента устранения неполноты и противоречивости с точки зрения использования ИНС. В рамках этого этапа осуществляется нейросетевая процедура устранения неполноты и противоречивости оценки и категоризации признаков СИО. Основу этого этапа составляет двухслойная ИНС. Сущность данного этапа заключается в том, что определяется хотя бы один признак, гарантированно включаемый в состав множества признаков нежелательной информации, причем этот признак однозначно включается в состав этого множества.

Искусственная нейронная сеть, применяемая в нашем случае, имеет традиционную двухслойную структуру и является классической нейронной сетью прямого распространения. Данный тип сети из-за своего широкого распространения в рамках тривиальных вычислений, выбран как простой и эффективный инструмент устранения неполноты и противоречивости оценки небольшого количества признаков нежелательной информации [41].

При этом число нейронов в слоях двухслойной искусственной нейронной сети соответствует количеству выбранных (анализируемых) признаков нежелательной информации, числу элементов вектора входных признаков, и может составлять от 1 до 50 нейронов для простой сети прямого распространения.

С помощью двухслойной искусственной нейронной сети формируется вектор входных признаков $\{\vec{C}_{\text{вх}}^{\,l}\}$, который учитывает неполные и противоречивые взаимосвязи всех признаков (по мнению L экспертов). По итогам решения задачи нейросетевого преобразования на выходе двухслойной ИНС получаем выходной вектор признаков СИО с коэффициентами (элементами), характеризующими вес (уровень опасности) этих признаков. Результаты этих вычислений позволяют (с учетом неполноты и противоречивости исходных данных) оценить и категорировать эту информацию, как нежелательную.

Предлагаемая модель выбора важных (значимых) признаков нежелательной информации в условиях неполноты и противоречивости позволяет избавиться от субъективных оценок и приобретать знания эмпирически, опираясь на мнения экспертов.

Пусть эмпирические данные имеют вид протокола:

$$\{\vec{C}_{\text{вх}}^{\,l}, \quad l = 1, \ldots, L\}, \tag{13}$$

где $\vec{C}^{l}_{\text{вх}} = (C^{l}_{\text{вх}\,1}, C^{l}_{\text{вх}\,2}, \ldots, C^{l}_{\text{вх}\,J})$ — вектор входных признаков (в терминах ИНС — входной вектор $\vec{A}$), который учитывает неполные и противоречивые взаимосвязи всех $j = 1, \ldots, J$ признаков нежелательной информации, по мнению l-го из множества L экспертов.

Показательным примером может служить вектор, характеризующий важность для каждого из пяти рассмотренных ранее признаков $\Delta\bar{j}_{\text{терр}}$, $\Delta\bar{j}_{\text{дет}}$, $\Delta\bar{j}_{\text{порн}}$, $\Delta\bar{j}_{\text{нарк}}$ и $\Delta\bar{j}_{\text{войн}}$:

$$\vec{A} = \vec{C}^{1}_{\text{вх}} = (1, 0, 0, 1, -1). \qquad (14)$$

Вектор (14) является символьной записью следующего выражения: «В соответствии с мнением первого эксперта первый признак СИО $C_{\text{вх}\,1}$ (имеет физический смысл $\Delta\bar{j}_{\text{терр}}$) и четвертый признак $C_{\text{вх}\,4}$ ($\Delta\bar{j}_{\text{нарк}}$) являются важными, существенными, значимыми. Пятый признак $C_{\text{вх}\,5}$ (имеет физический смысл $\Delta\bar{j}_{\text{войн}}$) является «не важным», не существенным. По остальным признакам ($C_{\text{вх}\,2}$ и $C_{\text{вх}\,3}$) мнение первого эксперта отсутствует».

Предположим, что в данный момент гарантированно важным, существенным и значимым признаком является признак $C_{\text{вх}\,5}$, характеризующий $\Delta\bar{j}_{\text{войн}}$. Другие признаки — неопределенны. Тогда в целях получения обоснованных результатов оценки смыслового содержания контента для поиска и обнаружения нежелательной информации, необходимо реконструировать недостающие компоненты вектора важных, существенных и значимых признаков.

Двухслойная ИНС реконструирует недостающие компоненты вектора $\vec{A}$. Рассмотрим этот процесс на примере. Предположим, нас интересуют составляющие вектора, характеризующего важность всех признаков при условии, что обязателен для включения в список опасных признаков именно пятый признак из всей совокупности признаков. Иными словами, значение $C_{\text{вх}\,5}$, характеризующее важность этого признака, равно «1». Нормируем приращения всех признаков относительно шкалы активационной функции. Пусть активационная функция имеет следующий ступенчатый вид:

$$f(C_{\text{вх}}) = \begin{cases} 1, & C_{\text{вх}} \geq 1; \\ 0, & 0 \leq C_{\text{вх}} < 1; \\ -1, & C_{\text{вх}} < 0. \end{cases} \quad (15)$$

Простая ступенчатая функция активации нейронов выбрана из множества возможных (ступенчатая, линейная, сигмоидальная, гиперболический тангенс, функция ReLu и др.) с учетом того, что в нашем, в сущности «бинарном», случае принятия решения о важности конкретного признака нежелательной информации для определения границы активации достаточно определить, превышает ли значение этого признака некоторое пороговое значение [27, 29]. В этом случае $C_{\text{вх }5}$, характеризующее важность признака $\Delta\bar{j}_{\text{войн}}$, будет соответствовать значению выхода 5-го нейрона, равному 1, а входной вектор примет вид $\vec{A}$ = (0, 0, 0, 0, 1). Тогда выходной вектор $\vec{B}$ = (b_1, b_2, b_3, b_4, b_5) двухслойной ИНС последовательно принимает следующие значения:

$\vec{B}$ (0) = f([0; 0; 0; 0; 1]) = [0, 0, 0, 0, 1];

$\vec{B}$ (1) = f([0,667; – 0,333; 1; 1; 0]) = [0, – 1, 1, 1, 1];

$\vec{B}$ (2) = f([3; – 0,667; 4; 4; 7]) = [1, – 1, 1, 1, 1];

$\vec{B}$ (3) = f([3; – 1,667; 4,667; 4,333; 7,667])= [1, – 1, 1, 1, 1];

$\vec{B}$ (4) = f([3; – 1,667; 4,667; 4,333; 7,667])= [1, – 1, 1, 1, 1];

$\vec{B}$ (5) = f([3; – 1,667; 4,667; 4,333; 7,667])= [1, – 1, 1, 1, 1].

Полученные результаты характеризуют суммарную предпочтительность включения данных признаков в состав множества опасных и могут быть представлены графически (рис. 9).

Из рисунка 9 видно, что двухслойная ИНС в интересах оценки смыслового содержания контента для поиска и обнаружения нежелательной информации стабилизировалась уже после третьего шага. Таким образом, с помощью такой ИНС, состоящей из двух слоев нейронов, можно осуществить оценку и краткосрочное нормативное прогнозирование веса (важности, значимости, опасности) признаков в условиях неполноты и противоречивости исходных данных. Результаты решения задачи в рамках второго примера позволяют с высокой степенью объективности, опираясь на накопленные в нейронной сети данные, сформировать вектор существенных признаков СИО и

корректно выбрать объем и номенклатуру признаков нежелательной информации. При этом следует заметить, что в состав множества опасных признаков гарантированно войдут такие признаки, как $C_{\text{вх }1}$ (имеет физический смысл $\Delta\bar{j}_{\text{терр}}$), $C_{\text{вх }3}$ ($\Delta\bar{j}_{\text{порн}}$), $C_{\text{вх }4}$ ($\Delta\bar{j}_{\text{нарк}}$) и $C_{\text{вх }5}$ (имеет физический смысл $\Delta\bar{j}_{\text{войн}}$), и не войдет признак $C_{\text{вх }2}$ ($\Delta\bar{j}_{\text{дет}}$).

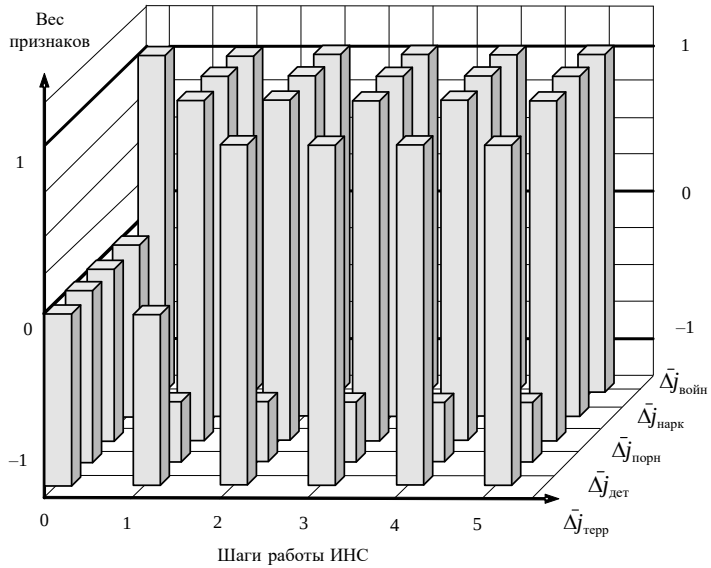


Рис. 9. Диаграмма зависимости веса признаков нежелательной информации от шага (цикла) вычисления новых состояний нейронов выходного слоя ИНС

Результаты реализации рассмотренных этапов работы компонента устранения неполноты и противоречивости результатов классификации СИО, а также результаты его экспериментальной оценки показывают, что использование обоих этапов в совокупности позволяет устранить неопределенность любого вида при формировании множества опасных, явных признаков для принятия решений в интересах выявления и противодействия нежелательной информации. Возможным очередным шагом исследований, нацеленным на совместное решение задач устранения как нечеткости, так и неполноты и противоречивости признаков нежелательной информации, может быть создание нейро-

нечетких математических моделей и алгоритмов обработки и интерпретации данных [33]. Данные механизмы, несмотря на сложность реализации, потенциально применимы, учитывая, что количество входов (признаков нежелательной информации) небольшое, а нечеткие алгоритмы и ИНС работают достаточно эффективно.

6. Заключение. В настоящей статье предложен новый тип интеллектуальных систем, ориентированный на аналитическую обработку цифрового сетевого контента в интересах защиты от нежелательной информации. Проведенный анализ состояния исследований в этой области показал, что автоматизированное обнаружение и противодействие нежелательной информации в цифровом сетевом контенте остается открытой проблемой. Предложенная архитектура ИСАОЦСК содержит три уровня, на которых располагаются компоненты распределенного сканирования сети, многоаспектной классификации сетевых ИО, устранения неполноты и противоречивости, принятия решений и визуализации. Рассмотрены модели и алгоритмы функционирования наиболее характерных компонентов системы, таких как компонент распределенного сканирования, компонент многоаспектной классификации, компонент устранения неполноты и противоречивости и компонент принятия решений. Для распределенных сетевых сканеров разработаны решения по реализации таких функций, как обнаружение и загрузка веб-контента, структурная категоризация, вычисление частотных характеристик и построение карт переходов СИО. В состав компонента многоаспектной классификации сетевых ИО предложено включить модули фильтрации, извлечения признаков, предобработки признаков и классификации СИО. При этом в функционировании этого компонента выделяются режим обучения и режим анализа. При обучении выполняется настройка классификаторов с помощью последовательностей обучающих векторов. При анализе определяется класс ИО, включая характер и степень вредоносности сетевого контента. Для устранения неполноты и противоречивости результатов классификации используются методы обработки нечетких знаний и обработка исходных данных с помощью ИНС. Принятие решений по противодействию нежелательной информации основывается на предложенных моделях информационной системы, СИО и контрмеры.

Экспериментальная оценка предложенных решений по построению и функционированию ИСАОЦСК показала, что предлагаемая система вполне отвечает предъявляемым к ней требованиям. Так, достоверность классификации нежелательных СИО в наборе данных, сформированном с помощью распределенных сетевых сканеров, достигала

84 процентов. Этот результат был получен в реальном масштабе времени при объеме набора данных, превышающем 8 Гбайт. Предложенные этапы работы компонента устранения неполноты и противоречивости результатов классификации СИО позволяют устранить в исходных данных для классификации СИО неопределенности любого вида в интересах принятия решений по выявления и противодействию нежелательной информации.

Направления дальнейших исследований связываются с усовершенствованием моделей и алгоритмов функционирования предложенной системы, расширяя область ее применения на обработку графического и мультимедийного веб-контента, а также обнаружение и противодействие недостоверной (фейковой) новостной информации.

Литература

1. *Scott J.* Social Network Analysis: Developments, Advances, and Prospects // Social Network Analysis and Mining. 2011. vol. 1. no. 1. pp. 21-26.
2. *Jebari C.* A pure URL-based genre classification of web pages // Proceedings of the 25th International Workshop on Database and Expert Systems Applications. 2014. pp. 233–237.
3. *Kotenko I., Chechulin A., Komashinsky D.* Categorisation of Web Pages for Protection against Inappropriate Content in the Internet // International Journal of Internet Protocol Technology (IJIPT). 2017. vol. 10. no. 1. pp. 61-71.
4. *Vaismoradi M., Turunen H., Bondas T.* Content Analysis and Thematic Analysis: Implications for Conducting a Qualitative Descriptive Study // Nursing & Health Sciences. 2013. vol. 15. no. 3. pp. 398-405.
5. *Defranco J.F., Laplante Ph.A.* A Content Analysis Process for Qualitative Software Engineering Research // Innov. Syst. Softw. Eng. 2017. vol. 13. no. 2-3. pp. 129-141.
6. *Boettger R.K., Palmer L.A.* Quantitative Content Analysis: Its Use in Technical Communication // IEEE Transactions on Professional Communication. 2010. vol. 53. no. 4. pp. 346-357.
7. *Linhares R.N., Costa A.P.* The use of qualitative data analysis software in brazilian educational papers // Proceedings of the International Conference in Engineering Applications (ICEA). 2019. pp. 1–7.
8. *Pashakhanlou H.* Fully Integrated Content Analysis in International Relations // International Relations. 2017. vol. 31. no. 4. pp. 447-465.
9. *Timmermans S., Iddo T.* Theory Construction in Qualitative Research: From Grounded Theory to Abductive Analysis // Sociological Theory. 2012. vol. 30. no. 3. pp. 167-186.
10. *Gunawan T.S., Abdullah N.A.J., Kartiwi M., Ihsanto E.* Social network analysis using python data mining // Proceedings of the 8th International Conference on Cyber and IT Service Management (CITSM). 2020. pp. 1–6.
11. UCINET documentation. URL: sites.google.com/site/ ucinetsoftware/document (дата доступа: 29.07.2021).
12. *Du W.* Toward semantic social network analysis for business big data // Proceedings of the 14th International Conference on Semantics, Knowledge and Grids (SKG). 2018. pp. 1–8.

13. *Li H., Zhang Z., Xu Y.* Web page classification method based on semantics and structure // Proceedings of the 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD). 2019. pp. 238–243.
14. *Patil A., Pawar B.* Automated classification of web sites using Naive Bayessian algorithm // Proceedings of the International Multi-Conference of Engineers and Computer Scientists. 2012. vol. 1. pp. 466–467.
15. *Kotenko I., Chechulin A., Shorov A., Komashinsky D.* Analysis and evaluation of web pages classification techniques for inappropriate content blocking // Proceedings of the 14th Industrial Conference on Data Mining (ICDM 2014). Lecture Notes in Artificial Intelligence. 2014. vol. 8557. pp. 39–54.
16. *Shibu S., Vishwakarma A., Bhargava N.* A Combination Approach for Web Page Classification using Page Rank and Feature Selection Technique // International Journal of Computer Theory and Engineering. 2010. vol. 2. no. 6. pp. 897-900.
17. *Xu Z., Yan F., Qin J., Zhu H.* A web page classification algorithm based on link information // Proceedings of the 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science. 2011. pp. 82–86.
18. *Hashemi M.* Web Page Classification: A Survey of Perspectives, Gaps, and Future Directions // Multimed. Tools Appl. 2020. vol. 79. pp. 11921-11945.
19. *Patel A.D., Pandya V.N.* Web page classification based on context to the content extraction of articles // Proceedings of the 2nd International Conference for Convergence in Technology (I2CT). 2017. pp. 539–541.
20. *Arya C., Dwivedi S.K.* News web page classification using URL content and structure attributes // Proceedings of the 2nd International Conference on Next Generation Computing Technologies (NGCT). 2016. pp. 317–322.
21. *Safae L., Habib B. E., Abderrahim T.* A Review of machine learning algorithms for web page classification // Proceedings of the 5th International Congress on Information Science and Technology (CiSt). 2018. pp. 220–226.
22. *Aydın K.E., Baday S.* Machine learning for web content classification // Proceedings of the Innovations in Intelligent Systems and Applications Conference (ASYU). 2020. pp. 1–7.
23. *Petprasit W., Jaiyen S.* E-commerce web page classification based on automatic content extraction // Proceedings of the 12th International Joint Conference on Computer Science and Software Engineering (JCSSE). 2015. pp. 74–77.
24. *Belmouhcine A., Idrissi A., Benkhalifa M.* Web Classification Approach Using Reduced Vector Representation Model Based on HTML Tags // Journal of Theoretical and Applied Information Technology. 2013. vol. 55. no. 1. pp. 137-148.
25. *Kotenko I., Chechulin A., Komashinsky D.* Evaluation of text classification techniques for inappropriate web content blocking // Proceedings of the IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS'2015). 2015. pp. 412–417.
26. *Novozhilov D., Kotenko I., Chechulin A.* Improving the categorization of web sites by analysis of html-tags statistics to block inappropriate content // Proceedings of the 9th International Symposium on Intelligent Distributed Computing (IDC'2015). 2016. pp. 257–263.
27. *Mishra M., Srivastava M.* A view of artificial neural network // Proceedings of the International Conference on Advances in Engineering & Technology Research (ICAETR - 2014). 2014, pp. 1–3.
28. *Mehlig B.* Artificial Neural Networks. University of Gothenburg, Sweden. 2019.
29. *Burghardt F., Garbe R.* Introduction of artificial neural networks in EMC // Proceedings of the IEEE Symposium on Electromagnetic Compatibility, Signal Integrity and Power Integrity (EMC, SI & PI). 2018. pp. 165–169.

30. *Parashchuk I.B.* System formation algorithm of communication network quality factors using artificial neural networks // Proceedings of the 1st IEEE International Conference on Circuits and System for Communications (ICCSC'02). 2002. pp. 263–266.
31. *Pandey K., Bhanacharjee S., Lau S., Tushir M.* A Comparative study of fuzzy systems and neural networks for system modeling and identification // Proceedings of the 2nd IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES). 2018. pp. 876–880.
32. *Агеев С.А., Саенко И.Б.* Управление безопасностью защищенных мультисервисных сетей специального назначения // Труды СПИИРАН. 2010. № 2(13). С. 182–198.
33. *Kotenko I., Parashchuk I., Omar T.* Neuro-fuzzy models in tasks of intelligent data processing for detection and counteraction of inappropriate, dubious and harmful information // Proceedings of the 2nd International Scientific-Practical Conference Fuzzy Technologies in the Industry. 2018. pp. 116–125.
34. *Нугуманова А.Б., Бессмертный И.А., Пецина П., Байбурин Е.М.* Обогащение модели Bag of words семантическими связями для повышения качества классификации текстов предметной области // Программные продукты и системы. 2016. №. 2 (114). С. 89–99.
35. *Mikolov T., Chen K., Corrado G., Dean J.* Efficient estimation of word representations in vector space // arXiv preprint arXiv:1301.3781. 2013. pp. 1–12.
36. SquidGuard – Blacklists. URL: www.squidguard.org/blacklists.html (дата доступа: 29.07.2021).
37. Shalla Secure Services. Shalla's Blacklists. URL: www.shallalist.de/ (дата доступа: 29.07.2021).
38. DMOZ. Archive. URL: dmoz-odp.org/ (дата доступа: 29.07.2021).
39. *Joulin A., Grave E., Bojanowski P., Mikolov T.* Bag of tricks for efficient text classification // arXiv preprint arXiv:1607.01759. 2016. pp. 1–5.
40. *Браницкий А.А., Котенко И.В.* Обнаружение сетевых атак на основе комплексирования нейронных, иммунных и нейронечетких классификаторов // Информационно-управляющие системы. 2015. № 4 (77). С. 69–77.
41. *Паращук И.Б., Башкирцев А.С., Михайличенко Н.В.* Анализ уровней и видов неопределенности, влияющей на принятие решений по управлению информационными системами // Информация и космос. 2017. № 1. С. 112–120.

Котенко Игорь Витальевич — д-р техн. наук, профессор, главный научный сотрудник, руководитель лаборатории, лаборатория проблем компьютерной безопасности, Федеральное государственное бюджетное учреждение науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН). Область научных интересов: безопасность компьютерных сетей, в том числе управление политиками безопасности, разграничение доступа, аутентификация, анализ защищенности, обнаружение компьютерных атак, межсетевые экраны, защита от вирусов и сетевых червей, анализ и верификация протоколов безопасности и систем защиты информации, защита программного обеспечения от взлома и управление цифровыми правами, технологии моделирования и визуализации для противодействия кибер-терроризму. Число научных публикаций — свыше 500. ivkote@comsec.spb.ru, http://www.comsec.spb.ru; 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ; р.т.: +7(812)328-3337, факс: +7(812)328-4450.

Саенко Игорь Борисович — д-р техн. наук, профессор, ведущий научный сотрудник, лаборатория проблем компьютерной безопасности, СПб ФИЦ РАН; профессор кафедры, Военная академия связи. Область научных интересов: автоматизированные информационные системы, информационная безопасность, обработка и передача данных по каналам связи, теория моделирования и математическая статистика, теория информации. Число

научных публикаций — свыше 400. ibsaen@comsec.spb.ru, http://www.comsec.spb.ru; 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ; р.т.: +7(812)328-3337, факс: +7(812)328-4450.

Браницкий Александр Александрович — к.т.н., старший научный сотрудник, лаборатория проблем компьютерной безопасности, СПб ФИЦ РАН. Область научных интересов: безопасность компьютерных сетей, искусственный интеллект, функциональное программирование. Число научных публикаций — 50. branitskiy@comsec.spb.ru, http://www.comsec.spb.ru; 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ; р.т.: +7(812)328-3337, факс: +7(812)328-4450.

Паращук Игорь Борисович — д-р техн. наук, профессор, ведущий научный сотрудник лаборатории проблем компьютерной безопасности, СПб ФИЦ РАН; профессор кафедры, Военная академия связи. Область научных интересов: анализ качества и эффективности автоматизированных информационных систем, центры обработки данных, безопасность информации, теория оценивания, теория моделирования и математическая статистика, теория информации, теория передачи данных. Число научных публикаций — свыше 250. parashchuk@comsec.spb.ru, http://www.comsec.spb.ru; 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ; р.т.: +7(812)328-3337, факс: +7(812)328-4450.

Гайфулина Диана Альбертовна — младший научный сотрудник, лаборатория проблем компьютерной безопасности, СПб ФИЦ РАН. Область научных интересов: безопасность компьютерных сетей, искусственный интеллект. Число научных публикаций — 20. gaifulina@comsec.spb.ru, http://www.comsec.spb.ru; 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ; р.т.: +7(812)328-3337, факс: +7(812)328-4450.

Поддержка исследований. Работа выполнена при финансовой поддержке РФФИ, проект № 18-29-22034 мк.

DOI 10.15622/ia.20.4.1

I. KOTENKO, I. SAENKO, A. BRANITSKIY,
I. PARASHCHUK, D. GAIFULINA

INTELLIGENT SYSTEM OF ANALYTICAL PROCESSING OF DIGITAL NETWORK CONTENT FOR HIS PROTECTION AGAINST INAPPROPRIATE INFORMATION

Kotenko I., Saenko I., Branitskiy A., Parashchuk I., Gaifulina D. **Intelligent System of Analytical Processing of Digital Network Content for His Protection Against Inappropriate Information.**

Abstract. Currently, the Internet and social networks as a medium for the distribution of digital network content are becoming one of the most important threats to personal, public and state information security. There is a need to protect the individual, society and the state from inappropriate information. In scientific and methodological terms, the problem of protection from inappropriate information has an extremely small number of solutions. This determines the relevance of the results presented in the article, aimed at developing an intelligent system of analytical processing of digital network content to protect against inappropriate information. The article discusses the conceptual foundations of building such a system, revealing the content of the concept of inappropriate information and representing the overall architecture of the system. Models and algorithms for the functioning of the most characteristic components of the system are given, such as a distributed network scanning component, a multidimensional classification component of network information objects, a component for eliminating incompleteness and inconsistency, and a decision-making component. The article presents the results of the implementation and experimental evaluation of system components, which demonstrated the ability of the system to meet the requirements for the completeness and accuracy of detection and counteraction of unwanted information in conditions of its incompleteness and inconsistency.

Keywords: Intelligent System, Digital Network Content, Inappropriate Information, Classification, Fuzzy Knowledge, Decision Making.

Kotenko Igor — Ph.D., Dr.Sci., Professor, Head of Laboratory, Laboratory of Computer Security Problems, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: computer network security, including security policy management, access control, authentication, network security analysis, intrusion detection, firewalls, deception systems, malware protection, verification of security systems, digital right management, modeling, simulation and visualization technologies for counteraction to cyber terrorism; The number of publications — over 500. ivkote@comsec.spb.ru, www.comsec.spb.ru; 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone: +7(812)328-3337, fax: +7(812)328-4450.

Saenko Igor — Ph.D., Dr.Sci., Professor; Leading research scientist, Laboratory of Computer Security Problems, SPC RAS; Professor of the department, the Military academy of communications. Research interests: automated information systems, information security, processing and transfer of data on data links, theory of modeling and mathematical statistics, information theory. The number of publications — over 400. ibsaen@comsec.spb.ru, www.comsec.spb.ru; 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone: +7(812)328-3337, fax: +7(812)328-4450.

Branitskiy Alexander — PhD, Senior researcher, Laboratory of Computer Security Problems, SPC RAS. Research interests: security of computer networks, artificial intelligence, functional

programming. The number of publications — 50. branitskiy@comsec.spb.ru, http://www.comsec.spb.ru; 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone: +7(812)328-3337, fax: +7(812)328-4450.

Parashchuk Igor — Ph.D., Dr.Sci., Professor; Leading research scientist, Laboratory of Computer Security Problems, SPC RAS; Professor of the department, the Military academy of communications. Research interests: analysis of the quality and efficiency of automated information systems, data processing centers, information security, estimation theory, modeling theory and mathematical statistics, information theory, data transmission theory. The number of publications — over 250. parashchuk@comsec.spb.ru, http://www.comsec.spb.ru; 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone: +7(812)328-3337, fax: +7(812)328-4450.

Gaifulina Diana — Junior researcher, Laboratory of Computer Security Problems, SPC RAS. Research interests: security of computer networks, artificial intelligence. number of publications — 20. gaifulina@comsec.spb.ru, http://www.comsec.spb.ru; 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone: +7(812)328-3337, fax: +7(812)328-4450.

Acknowledgements. This research was funded by RFBR according to the research project No. 18-29-22034 mk.

References

1. Scott J. Social Network Analysis: Developments, Advances, and Prospects. *Social Network Analysis and Mining*. 2011. vol. 1. no. 1. pp. 21-26.
2. Jebari C. A pure URL-based genre classification of web pages. Proceedings of the 25th International Workshop on Database and Expert Systems Applications. 2014. pp. 233–237.
3. Kotenko I., Chechulin A., Komashinsky D. Categorisation of Web Pages for Protection Against Inappropriate Content in the Internet. *International Journal of Internet Protocol Technology (IJIPT)*. 2017. vol. 10. no. 1. pp. 61-71.
4. Vaismoradi M., Turunen H., Bondas T. Content Analysis and Thematic Analysis: Implications for Conducting a Qualitative Descriptive Study. *Nursing & Health Sciences*. 2013. vol. 15. no. 3. pp. 398-405.
5. Defranco J.F., Laplante Ph.A. A Content Analysis Process for Qualitative Software Engineering Research. *Innov. Syst. Softw. Eng*. 2017. vol. 13. no. 2–3. pp. 129-141.
6. Boettger R.K., Palmer L.A. Quantitative Content Analysis: Its Use in Technical Communication. *IEEE Transactions on Professional Communication*. 2010. vol. 53. no. 4. pp. 346-357.
7. Linhares R.N., Costa A.P. The use of Qualitative Data Analysis Software In Brazilian Educational Papers. Proceedings of the International Conference in Engineering Applications (ICEA). 2019. pp. 1–7.
8. Pashakhanlou H. Fully Integrated Content Analysis in International Relations. *International Relations*. 2017. vol. 31. no. 4. pp. 447–465.
9. Timmermans S., Iddo T. Theory Construction in Qualitative Research: From Grounded Theory to Abductive Analysis. *Sociological Theory*. 2012. vol. 30. no. 3. pp. 167-186.
10. Gunawan T.S., Abdullah N.A.J., Kartiwi M., Ihsanto E. Social Network Analysis using Python Data Mining. Proceedings of the 8th International Conference on Cyber and IT Service Management (CITSM), 2020, pp. 1–6.
11. UCINET documentation. Available at: sites.google.com/site/ucinetsoftware/document (accessed: 29.07.2021).

12. Du W. Toward semantic social network analysis for business big data. Proceedings of the 14th International Conference on Semantics, Knowledge and Grids (SKG). 2018. pp. 1–8.
13. Li H., Zhang Z., Xu Y. Web page classification method based on semantics and structure. Proceedings of the 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD). 2019. pp. 238–243.
14. Patil A., Pawar B. Automated classification of web sites using Naive Bayessian algorithm. Proceedings of the International Multi-Conference of Engineers and Computer Scientists. 2012. vol. 1. pp. 466–467.
15. Kotenko I., Chechulin A., Shorov A., Komashinsky D. Analysis and evaluation of web pages classification techniques for inappropriate content blocking. Proceedings of the 14th Industrial Conference on Data Mining (ICDM 2014). Lecture Notes in Artificial Intelligence. 2014. vol. 8557. pp. 39–54.
16. Shibu S., Vishwakarma A., Bhargava N. A Combination Approach for Web Page Classification Using Page Rank and Feature Selection Technique. *International Journal of Computer Theory and Engineering*. 2010. vol. 2. no. 6. pp. 897-900.
17. Xu Z., Yan F., Qin J., Zhu H. A web page classification algorithm based on link information. Proceedings of the 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science. 2011. pp. 82–86.
18. Hashemi M. Web Page Classification: A Survey of Perspectives, Gaps, and Future Directions. *Multimed. Tools Appl.* 2020. vol. 79. pp. 11921-11945.
19. Patel A.D., Pandya V.N. Web page classification based on context to the content extraction of articles. Proceedings of the 2nd International Conference for Convergence in Technology (I2CT). 2017. pp. 539–541.
20. Arya C., Dwivedi S.K. News web page classification using URL content and structure attributes. Proceedings of the 2nd International Conference on Next Generation Computing Technologies (NGCT). 2016. pp. 317–322.
21. Safae L., Habib B. E., Abderrahim T. A review of machine learning algorithms for web page classification. Proceedings of the 5th International Congress on Information Science and Technology (CiSt). 2018. pp. 220–226.
22. Aydın K.E., Baday S. Machine learning for web content classification. Proceedings of the Innovations in Intelligent Systems and Applications Conference (ASYU). 2020. pp. 1–7.
23. Petprasit W., Jaiyen S. E-commerce web page classification based on automatic content extraction. Proceedings of the 12th International Joint Conference on Computer Science and Software Engineering (JCSSE). 2015. pp. 74–77.
24. Belmouhcine A., Idrissi A., Benkhalifa M. Web Classification Approach Using Reduced Vector Representation Model Based on HTML Tags. *Journal of Theoretical and Applied Information Technology*. 2013. vol. 55. no. 1. pp. 137-148.
25. Kotenko I., Chechulin A., Komashinsky D. Evaluation of text classification techniques for inappropriate web content blocking. Proceedings of the IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS'2015). 2015. pp. 412–417.
26. Novozhilov D., Kotenko I., Chechulin A. Improving the categorization of web sites by analysis of html-tags statistics to block inappropriate content. Proceedings of the 9th International Symposium on Intelligent Distributed Computing (IDC'2015). 2016. pp. 257–263.
27. Mishra M., Srivastava M. A view of artificial neural network. Proceedings of the International Conference on Advances in Engineering & Technology Research (ICAETR - 2014). 2014, pp. 1–3.
28. Mehlig B. Artificial Neural Networks. University of Gothenburg, Sweden. 2019.

29. Burghardt F., Garbe R. Introduction of artificial neural networks in EMC. Proceedings of the IEEE Symposium on Electromagnetic Compatibility, Signal Integrity and Power Integrity (EMC, SI & PI). 2018. pp. 165–169.
30. Parashchuk I.B. System formation algorithm of communication network quality factors using artificial neural networks. Proceedings of the 1st IEEE International Conference on Circuits and System for Communications (ICCSC'02). 2002. pp. 263–266.
31. Pandey K., Bhanacharjee S., Lau S., Tushir M. A comparative study of fuzzy systems and neural networks for system modeling and identification. Proceedings of the 2nd IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES). 2018. pp. 876–880.
32. Ageev S.A., Saenko I.B. [Security management of protected multi-service networks for special purposes]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2010. № 2(13). pp. 182–198. (In Russ.).
33. Kotenko I., Parashchuk I., Omar T. Neuro-fuzzy models in tasks of intelligent data processing for detection and counteraction of inappropriate, dubious and harmful information. Proceedings of the 2nd International Scientific-Practical Conference Fuzzy Technologies in the Industry. 2018. pp. 116–125.
34. Nugumanova A.B., Bessmertny I.A., Petsina P., Bayburin E.M. [Enriching the Bag of words model with semantic links to improve the quality of classification of domain texts]. *Programmnye produkty i sistemy – Software products and systems*. 2016. № 2 (114). pp. 89–99. (In Russ.).
35. Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space. *arXiv preprint* arXiv:1301.3781. 2013. pp. 1–12.
36. SquidGuard – Blacklists. Available at: www.squidguard.org/blacklists.html (accessed: 29.07.2021).
37. Shalla Secure Services. Shalla's Blacklists. Available at: www.shallalist.de/ (accessed: 29.07.2021).
38. DMOZ. Archive. Available at: dmoz-odp.org/ (accessed: 29.07.2021).
39. Joulin A., Grave E., Bojanowski P., Mikolov T. Bag of tricks for efficient text classification. *arXiv preprint* arXiv:1607.01759. 2016. pp. 1–5.
40. Branitskiy A.A., Kotenko I.V. [Network attack detection based on combination of neural, immune and neuro-fuzzy classifiers]. *Informacionno-upravljajushhie sistemy – Information management systems*. 2015. № 4 (77). pp. 69–77. (In Russ.).
41. Parashchuk I.B., Bashkircev A.S., Mihajlichenko N.V. [Analysis of the levels and types of uncertainty affecting decision-making in the management of information systems]. *Informacija i kosmos – Information and space*. 2017. № 1. pp. 112–120. (In Russ.).

УДК 004.75 **DOI** 10.15622/ia.20.4.2

И.М. Шилов, Д.А. Заколдаев

БЕЗОПАСНОСТЬ ПРОТОКОЛА ПОИСКА И ВЕРИФИКАЦИИ В МНОГОМЕРНОМ БЛОКЧЕЙНЕ

Шилов И.М., Заколдаев Д.А. **Безопасность протокола поиска и верификации в многомерном блокчейне.**

Аннотация. Проблема безопасного обмена информацией и проведения транзакций между устойчивыми распределенными реестрами является одной из наиболее актуальных в сфере проектирования и построения децентрализованных технологий. До настоящего времени были предложены подходы, ориентированные на ускорение проверки цепочки блоков для верификации транзакций в соседних блокчейнах. При этом проблема поиска ранее не затрагивалась. В работе рассмотрен вопрос безопасности обмена данными между самостоятельными устойчивыми распределенными реестрами в рамках многомерного блокчейна. Описаны принципы и основные этапы работы протокола, а также базовые требования, предъявляемые к нему. Предложены способы построения протокола обмена сообщениями для верификации внешних транзакций: централизованный подход, принцип подмножества и стойкий SVP. Доказана эквивалентность централизованного подхода идеальному функционалу поиска и верификации в GUC-моделях. Показана вероятность успешной верификации в случае использования подхода, основанного на подмножествах, при применении полного графа сети или эквивалентного подхода с полным графом между родительским и дочерним блокчейнами. Доказана небезопасность случая со связью 1 к 1 между родительским и дочерним реестром, а также небезопасность подхода, основанного на подмножестве узлов родительского и дочернего реестров. Предложен стойкий протокол поиска и верификации блоков и транзакций, основанный на свойствах стойкости устойчивых распределенных реестров. В значительной степени вероятность атаки определяется вероятностью атаки на процесс верификации, а не на процесс поиска. При необходимости защиты от атакующих, контролирующих до половины узлов в сети, предложен метод комбинации подходов для поиска и верификации блоков и транзакций.

Ключевые слова: протокол поиска и верификации, блокчейн, сайдчейн, многомерный блокчейн, GUC-фреймворк, устойчивый распределенный реестр.

1. Введение.

Многомерный блокчейн представляет собой один из подходов к построению устойчивого распределенного реестра. Эта технология основана на принципах работы обычного одномерного блокчейна и призвана решить основные проблемы работы одномерного блокчейна с сохранением гарантий безопасности и основных метрик его работы [1,2]. Наиболее существенными среди решаемых проблем являются проблема масштабирования устойчивых распределенных реестров и проблема безопасного обмена информацией между устойчивыми распределенными реестрами [3-5].

Внешние транзакции в многомерном блокчейне проводятся в два этапа (инициирования и акцепта), которые выполняются узлами в создавшем и принимающем транзакцию реестрах соответственно. Для успешной проверки существования и корректности транзакции в произвольном реестре предполагается использование специального

протокола. Он предназначен для достижения трех целей: поиска узлов, поддерживающих инициирующий реестр, запроса у них информации о корректности транзакции и принятия решения о допустимости принятия транзакции (выполнения фазы акцепта).

Представленный в предыдущих работах анализ многомерного блокчейна был посвящен технологии и ее безопасности в целом, без рассмотрения особенностей функционирования протокола поиска и верификации блоков и транзакций [1,6].

В работе [1] была сформирована концепция многомерного блокчейна, его структура и некоторые принципы функционирования (например, адресация). Были продемонстрированы его достоинства в сопоставлении с существующими аналогами, а также обобщенно показан подход к проведению внешних транзакций. При этом явным образом не рассматривалась безопасность технологии и подходы к организации взаимодействия между реестрами.

В работе [6] была построена модель устойчивого распределенного реестра с использованием фреймворка универсальной композиции (GUC, Generalized Universal Composability framework), GUC-модель, основанная на существующих моделях устойчивых распределенных реестров. Она предназначена для доказательства безопасности многомерного блокчейна. В работе явным образом введено понятие протокола поиска и верификации блоков и транзакций, однако в GUC-модели вместо реализаций данного протокола используется идеальный функционал, работающий по принципу черного ящика и предоставляющий необходимые функции узлам. Была доказана безопасность устойчивого распределенного реестра на основе многомерного блокчейна, построенного с использованием данной модели устойчивого распределенного реестра и идеального функционала проведения внешних транзакций. При этом возможные способы реализации протокола рассмотрены не были.

Корректное функционирование протокола поиска и верификации является основным условием корректности существующих утверждений, касающихся безопасности многомерного блокчейна. Поэтому актуальной представляется задача проектирования такого протокола, а также формального доказательства его безопасности, то есть сохранения свойств устойчивых распределенных реестров при его использовании.

В данной работе рассмотрен вопрос обмена данными между устойчивыми распределенными реестрами в пределах многомерного блокчейна. Авторами впервые предложены несколько подходов для организации такого обмена, а также произведена оценка их

безопасности. На основе полученных результатов предложен устойчивый протокол поиска и верификации блоков и транзакций, безопасность которого также проанализирована. Полученные результаты могут быть использованы как для реализации устойчивых распределенных реестров в пределах многомерного блокчейна, так и для организации взаимодействия между самостоятельными устойчивыми распределенными реестрами.

Работа состоит из восьми разделов. В начале кратко проанализирована проблема межсистемного обмена и предложенные ранее способы ее решения. Далее сформулированы базовые требования к протоколу поиска и верификации и рассмотрен порядок его работы. Затем рассмотрены различные подходы к построению протокола поиска и верификации блоков и транзакций, проанализирована их безопасность. В заключении произведена оценка полученных результатов и указаны перспективы для дальнейших исследований по тематике.

2. Анализ проблемы межсистемного обмена в распределенных технологиях. Понятие устойчивого распределенного реестра возникло сравнительно недавно, практически одновременно с возникновением понятия «блокчейн». Хотя методы решения задачи Византийских генералов рассматривались и ранее, существенных успехов удалось достичь лишь во втором десятилетии XXI века. Анализу безопасности устойчивых распределенных реестров и технологии блокчейн посвящено множество работ. Некоторые рассматривают безопасность конкретных систем (например, [2]), другие посвящены анализу безопасности механизмов достижения консенсуса [7-10]. Вопросы межсистемного взаимодействия изучены в меньшей степени, хотя проблема обмена информацией между устойчивыми распределенными реестрами неоднократно рассматривалась в научных публикациях.

Наиболее простым подходом является отслеживание всей цепочки блоков при верификации внешней транзакции [11]. Существенным недостатком данного подхода является скорость принятия решения о корректности транзакции: требуется запрос и проверка всей цепочки блоков. Одним из первых подходов к ускорению этого процесса является использование вложенных блокчейнов (interlink) [11]. Вместо проверки всей цепочки блоков проверяется цепочка блоков с хэш-суммами менее, чем $T / 2^i$ (T – целевое значение хэш-суммы для механизма достижения консенсуса). В результате сложность алгоритма проверки значительно уменьшается.

В работе [12] данный подход был усовершенствован. Предложенное решение упрощает проверку внешних транзакций, причем алгоритм имеет логарифмическую сложность относительно длины цепочки блоков в проверяемом блокчейне. Основным преимуществом является возможность верификации транзакции с использованием только одного запроса к целевому реестру. При этом доказывается уязвимость алгоритма из [11] к атаке со стороны участника системы, обладающего менее чем 50% вычислительной мощности. Также введены дополнительные предикаты, которые обобщают концепцию верификации, введенную ранее. Их особенностью является параметризация зависящим от конкретной реализации оператором сопоставления предоставленных доказательств, что делает решение независимым от конкретной реализации и особенностей конкретной системы. Тем не менее, стоит отметить, что данное решение подходит только для блокчейнов, использующих доказательство работы в качестве механизма достижения консенсуса.

В работе [13] рассмотрен подход, когда сайдчейны используются исключительно для создания усовершенствованных операций над токенами без изменения принципов функционирования блокчейна. Приведено интуитивное доказательство безопасности.

В [14] представлен обобщенный подход к построению сайдчейнов. Особенностью работы является ее направленность на системы, использующие доказательства доли владения. Кроме того, этот подход позволяет строить сайдчейны, поддерживающие метод GHOST [15]. Предложен новый криптографический примитив, направленный на проверку существования транзакции в сайдчейне. Отличительной чертой работы является формализация понятия сайдчейна без связи с конкретным механизмом достижения консенсуса.

Обзор многих современных технологий для построения привязанных сайдчейнов (pegged sidechains) приведен в [16]. Выделяются следующие способы организации сайдчейнов: централизованный посредник, федеративные сайдчейны, SPV (Simple Payment Verification). Описанные в работе решения основаны на концепции криптовалют, допускается лишь перевод части токенов между системами путем их заморозки в одной и создания в другой цепочке блоков. Также стоит отметить, что все рассмотренные в этой работе решения представляют собой лишь практические реализации.

Большинство рассмотренных методов посвящены обмену информацией о платежах в приложениях, реализующих криптовалюты. С одной стороны, этот подход позволяет узлам не отслеживать сторонние цепочки блоков, поскольку верификация основана на

методах криптографии и не подразумевает принятие решения на основе суждений узлов, поддерживающих внешний блокчейн, о корректности транзакции. С другой стороны, эти методы не могут быть использованы в сферах, отличных от криптовалют.

Еще одним недостатком сайдчейнов является необходимость использования общей формы хранения информации о транзакциях. В противном случае узлы обязаны обладать информацией о способе интерпретации структуры блоков в блокчейне-инициаторе транзакции. Поэтому существующие решения не подходят для построения взаимодействия между системами, реализующими разные приложения на основе устойчивых распределенных реестров. Иными словами, проблема заключается в том, что безопасность достигается проверкой цепочки блоков в соседнем реестре и не учитывает ответы поддерживающих его узлов.

Наконец, существенным недостатком является ограничение взаимодействия двумя реестрами. Поэтому важной особенностью перечисленных работ является отсутствие поиска узлов, поддерживающих внешний блокчейн. Обычно адреса для организации обмена предоставляются алгоритму в качестве исходных данных.

Рассмотренные проблемы существующих решений и работ, посвященных анализу их безопасности, подтверждают актуальность задачи и практическую ценность данной работы.

3. Протокол поиска и верификации. Основное назначение протокола поиска и верификации заключается в сохранении свойств стойкости и живости для самостоятельных реестров при использовании многомерного блокчейна [17]. Стойкость (persistence) может быть нарушена только при неправильном подборе параметров проверки внешних транзакций. В этом случае входящая транзакция, включенная в момент, когда она не перешла в неизменное состояние в исходном реестре, может быть исключена из исходного реестра после регистрации в реестре-приемнике. Узлы, поддерживающие целевой реестр, не смогут получить информацию об этом событии при отсутствии соответствующих механизмов в протоколе поиска и верификации блоков и транзакций. Поэтому либо должен быть включен такой функционал, либо протокол поиска и верификации должен явным образом предотвращать подобное поведение. Живость (liveness) может быть нарушена в том случае, если узлы, поддерживающие принимающий транзакцию блокчейн, не получат подтверждение корректности корректной транзакции. В этом случае применение транзакции к реестру не произойдет за конечное время.

Еще одно существенное требование, предъявляемое к протоколу поиска и верификации – сохранение децентрализованной структуры решения. Основное преимущество блокчейн решений – возможность достижения взаимопонимания между самостоятельными узлами, действующими в ненадежной среде передачи информации в присутствии атакующих. При верификации внешних транзакций этот принцип должен сохраняться, поскольку использование централизованного аналога сводит на нет получаемые от использования децентрализации преимущества.

Работа протокола поиска и верификации блоков и транзакций включает несколько процедур:

1. Процедура инициализации (initialization).
2. Процедура поддержания работы сети (maintenance).
3. Процедура поиска (search).
4. Процедура верификации (verification).

В рамках процедуры инициализации происходит настройка основных параметров функционирования протокола, устанавливаются необходимые соединения с узлами, поддерживающими другие реестры. Эту процедуру каждый узел выполняет каждый раз при запуске и начале работы с функционалом внешних транзакций. Основными параметрами данной подпрограммы являются: множество сетевых адресов соседних узлов для организации взаимодействия, минимально и максимально допустимое количество узлов при активном взаимодействии, количество узлов в родительском и дочерних реестрах для организации взаимодействия, а также флаг, определяющий участие узла в поддержании работы многомерного блокчейна. Кроме того, важным параметром является количество блоков, необходимых для приведения транзакции в необратимый вид. Этот параметр определяется для каждого блокчейна.

Процедура поддержания работы сети заключается в сохранении установленных соединений с узлами, поддерживающими различные блокчейны в пределах многомерного блокчейна, установлении новых соединений при нарушении работоспособности существующих, а также периодической ротации этих соединений. Данная процедура настраивается временными параметрами, определяющими период обмена сообщениями для поддержания соединений и период изменения набора взаимодействующих соседних узлов.

Процедура поиска выполняется каждый раз, когда требуется проверка внешней транзакции. Путем выполнения запросов к другим узлам, поддерживающим многомерный блокчейн, осуществляется поиск узлов, поддерживающих реестр-инициатор внешней транзакции.

Параметры для данной процедуры настроены на этапе инициализации. В процедуре верификации внешней транзакции осуществляется выбор узлов для запроса информации, получение информации и принятие на основе этой информации решения о корректности или некорректности входящей внешней транзакции. Параметром является требуемая доля утвердительных ответов, полученных от опрашиваемых узлов, для принятия внешней транзакции. Работа протокола поиска и верификации внешних транзакций продемонстрирована на рисунке 1.

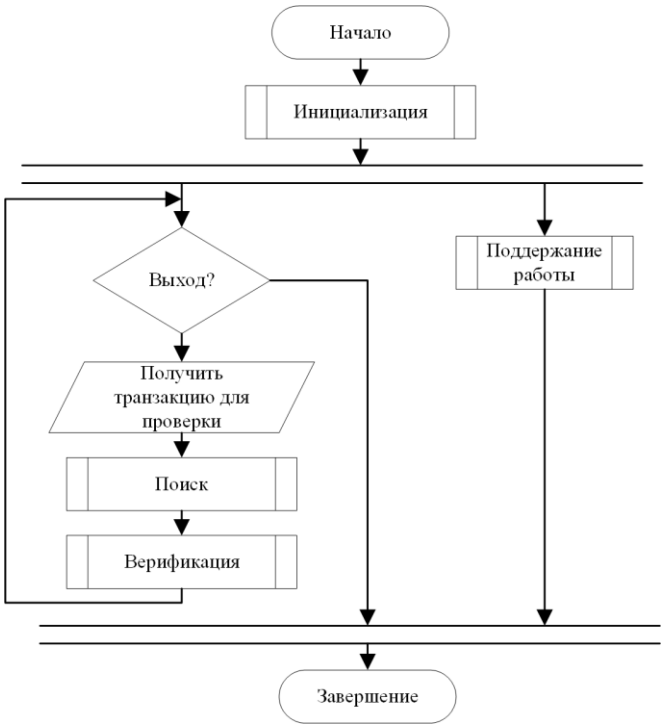


Рис. 1. Этапы работы протокола поиска и верификации

Возможны несколько подходов к построению протокола поиска и верификации блоков и транзакций:

1. Централизованный подход – проверка транзакций осуществляется с использованием единого узла или небольшой группы узлов, выступающих посредниками при проведении транзакций.

2. Подход, основанный на подмножествах – каждый узел поддерживает связь с некоторым подмножеством узлов в своем родительском блокчейне.
3. Стойкий search and verification protocol (SVP) – подход, основанный на свойствах устойчивого реестра и составляющий основной результат данной работы.

Рассмотрим предложенные подходы, выделим основные достоинства и недостатки этих подходов, а также произведем оценку безопасности каждого подхода.

3. Централизованный подход. Централизованный подход к построению протокола поиска и верификации подразумевает использование одного или нескольких узлов, совместно выполняющих работу по верификации транзакций. Они отслеживают состояние реестров в пределах многомерного блокчейна и предоставляют информацию о существовании или отсутствии транзакций при получении соответствующего запроса. Подобный подход используется в системе Hyperledger Fabric при создании и упорядочивании блоков: в качестве механизма достижения консенсуса используется получение подписей блоков со стороны удостоверяющих узлов.

Централизованный подход допустим в случаях, когда используются централизованные механизмы достижения консенсуса. Кроме того, иногда такой подход необходим, если транзакции между устойчивыми распределенными реестрами целенаправленно отслеживаются и анализируются контролирующими органами, и живость транзакций не является критически важным требованием. Стоит отметить, что аналогичных результатов можно добиться и в полноценно децентрализованном решении. Еще одним примером условий, при которых допустимо использование централизованного протокола поиска и верификации является ненагруженная система, в которой транзакции в каждом блокчейне сравнительно редки, и поддержание постоянных соединений для проведения внешних транзакций является избыточным.

Основным достоинством централизованного подхода является простота реализации, а также возможность выборочного блокирования транзакций в случае необходимости. Недостатками можно считать фактическое появление посредника при межсистемном обмене, что сводит на нет преимущества, предлагаемые многомерным блокчейном, а также достаточно низкую пропускную способность, поскольку количество запросов в единицу времени, которые может обработать единый узел, ограничено, что может привести к появлению дополнительной точки отказа.

Утверждение 1. Централизованный протокол поиска и верификации транзакций эквивалентен идеальному функционалу при условии, что узел, поддерживающий протокол честный.

Под честным узлом понимается узел, который не нарушает правила работы исполняемого протокола. В случае протокола поиска и верификации честность означает отсутствие действий, направленных на подтверждение несуществующих или отказ в подтверждении существующих транзакций. В этом случае централизованный узел поддерживает базу транзакций и отвечает на запросы по требованию, что полностью соответствует описанию идеального функционала, реализующего протокол поиска и верификации в GUC-моделях [6,18-21]. Следовательно, централизованный протокол UC-реализует протокол поиска и верификации при условии истинности предположений о его функционировании.

4. Подход, основанный на подмножествах. Иным способом работы с протоколом поиска и верификации является использование множеств. В этом случае каждый узел поддерживает взаимодействие с определенным подмножеством узлов, участвующих в работе многомерного блокчейна. Возможны четыре способа построения протокола поиска и верификации, основанного на подмножествах:

1. Способ, основанный на полном графе, когда каждый узел поддерживает соединение со всеми узлами системы.
2. Способ, основанный на полном графе, когда каждый узел поддерживает соединение со всеми узлами родительского реестра.
3. Способ, основанный на связи 1 к 1, когда каждый узел поддерживает связь только с одним узлом родительского реестра.
4. Способ, основанный на связи N к N, когда каждый узел поддерживает связь с набором узлов родительского блокчейна.

Второй и третий способы представляют собой граничные случаи подхода, основанного на подмножествах при взаимодействии только с родительским реестром, тогда как последний – промежуточный вариант. Рассмотрим каждый из описанных подходов. Далее под «соседними» реестрами понимаются реестры, находящиеся в отношении «родитель-дочерний реестр».

4.1. Способ полного графа сети со всеми узлами многомерного блокчейна. Способ, при котором узел поддерживает соединение со всеми узлами многомерного блокчейна, представляет собой вариант полного графа сетевого взаимодействия. Данный подход теоретически может применяться в сетях с небольшим количеством

узлов и редкими транзакциями, хотя его преимущества перед централизованным вариантом остаются предметом дискуссии. Фактически затраты на поддержание межсистемного взаимодействия могут превосходить затраты по поддержанию собственного реестра.

При проверке внешней транзакции узлы реестра-акцептора запрашивают информацию о существовании и корректности транзакции у всех узлов, поддерживающих реестр-инициатор, или у подмножества из k узлов. Поскольку стойкость реестра-инициатора постулируется по предположению, по принципу большинства верификация внешней транзакции гарантированно завершается успешно (поскольку честных узлов в стойком реестре большинство). При использовании подхода с запросом k узлов вероятность корректной верификации:

$$P = \begin{cases} \sum_{i=0}^{\left\lceil \frac{k}{2} \right\rceil - 1} C_k^i * q^i * p^{k-i}, \text{ если } z = \left\lceil \frac{k}{2} \right\rceil - 1 \le N_A; \\ 1, \text{ если } \left\lceil \frac{k}{2} \right\rceil - 1 > N_A, \end{cases} \quad (1)$$

где k – мощность подмножества, z – количество атакующих в подмножестве, N_A – общее число атакующих, P – целевое значение вероятности. В формуле высчитывается вероятность того, что не более половины из выбранных k узлов окажутся атакующими. Для этого суммируются все соответствующие вероятности, полученные по формуле Бернулли.

График изменения данной величины в зависимости от количества опрашиваемых узлов (k) приведен на рисунке 2. В качестве исходных величин использовались: количество атакующих 20, количество узлов 100, то есть атакующие исходно составляют 20% от общего числа узлов. С увеличением мощности множества вероятность успешной верификации стремится к единице. Поэтому при использовании данного подхода возможно достижение статистически безопасного обмена информацией о транзакциях даже при взаимодействии с подмножеством узлов.

Основные достоинства и недостатки этого подхода практически полностью повторяют достоинства и недостатки централизованного подхода. Хотя реализация такого подхода представляется задачей более сложной, чем реализация централизованного решения, пропускная способность в общем случае оказывается выше за счет параллельной обработки запросов.

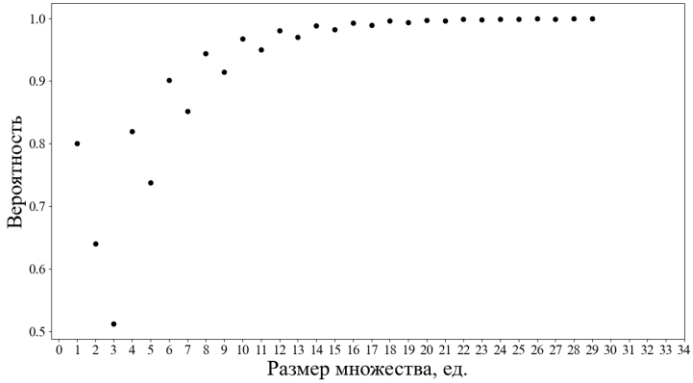


Рис. 2. Вероятность корректной верификации транзакций в полном графе

Утверждение 2. Протокол поиска и верификации блоков и транзакций, построенный на основе полного графа сетевого взаимодействия UC-реализует идеальный протокол поиска и верификации блоков и транзакций с вероятностью, указанной в Соотношении 1.

Поскольку Соотношение 1 выражает вероятность появления более половины честных узлов в выборке для запроса информации, оно отражает и вероятность принятия корректного решения по принципу большинства, что при исполнении прочих условий означает корректность реализации идеального протокола поиска и верификации. Результаты расчетов для $N = 100$ приведены в таблице 1.

Таблица 1. Вероятность успешной верификации

N	k	$q_1 = 0,5$	$q_2 = 0,4$	$q_3 = 0,3$	$q_4 = 0,2$	$q_5 = 0,1$
100	10	0,1719	0,3823	0,6496	0,8791	0,9872
100	20	0,2517	0,5956	0,8867	0,99	0,9999
100	30	0,2923	0,7145	0,9599	0,9991	1
100	40	0,3179	0,7911	0,9852	0,9999	1
100	50	0,3359	0,8438	0,9944	1	1

4.2. Способ полного графа между узлами соседних реестров. При использовании полного графа с родительским блокчейном необходимо рассмотреть безопасность как протокола верификации, так и протокола поиска. При поиске узлов целевого блокчейна для взаимодействия каждый атакующий узел может передавать

произвольное количество сгенерированных им узлов. В этом случае отсутствует контроль за количеством узлов, и атакующий может обеспечить себе сколь угодно большое преимущество перед честными узлами.

Поскольку анализ заведомо небезопасной реализации не имеет смысла, предлагается подход, защищающий систему от подобной атаки. Пусть при протоколе поиска на каждом уровне узлы получают от каждого узла перечень узлов его родительского или дочернего блокчейна, с которыми установлено соединение. Все узлы каждого реестра поддерживают соединение со всеми узлами соседнего реестра, поэтому для честных узлов предоставленные списки будут совпадать. Поскольку атакующий не контролирует большинство узлов, достаточно установить соединение с теми узлами, которые были найдены в более чем половине из предоставленных списков. Тогда в окончательном множестве по принципу большинства окажутся только узлы, с которыми установлено соединение честных узлов, то есть все узлы следующего блокчейна. Все остальные этапы взаимодействия эквивалентны случаю с полным графом и соединением «все-со-всеми».

4.3. Способ связи 1 к 1 между узлами соседних реестров. Это другой пограничный способ при построении взаимодействия с родительским реестром без учета особенностей его работы. Узлы для взаимодействия выбираются случайно. При этом возможно два варианта: либо узел, с которым установлено соединение, меняется в каждом раунде, либо соединения постоянны. Рассмотрим работу протокола в отдельно взятом раунде, поскольку в пределах одного раунда эти варианты идентичны, а вероятность выбора атакующего в наихудшем варианте постоянна.

Основным параметром в данной модели является расстояние между реестрами. Протокол поиска подразумевает последовательное взаимодействие с узлами всех реестров до корневого и далее от корневого до целевого реестра. Обозначим общее количество шагов d. На каждом уровне вероятность взаимодействия с атакующим совпадает с долей атакующих в реестре, с которым осуществляется взаимодействие – p_i. Если на каком-то промежуточном уровне начинается взаимодействие с атакующим, он может возвращать только подконтрольные атакующему узлы в следующих реестрах. Поэтому единственно безопасным исходом является взаимодействие с честными узлами на каждом уровне. Вероятность атаки на протокол поиска и верификации:

$$P = 1 - \prod_{i=1}^{d} (1 - p_i), \quad (2)$$

где p_i – вероятность взаимодействия с атакующим, d – число промежуточных блокчейнов. График вероятности атаки на протокол в зависимости от дистанции (для различных значений вероятности взаимодействия с атакующим на каждом шаге) приведен на рисунке 3.

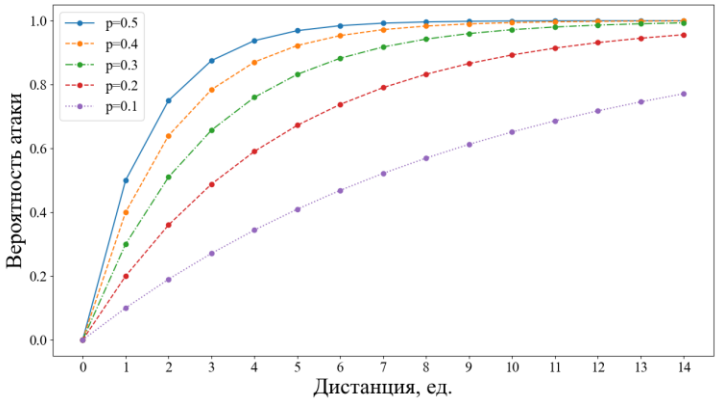


Рис. 3. Вероятность атаки на протокол в случае связи 1 к 1

Даже при незначительном количестве атакующих в каждом последующем реестре с увеличением количества промежуточных реестров вероятность атаки на протокол стремится к 1. Даже в случае, если количество атакующих составляет не более 10% в каждом реестре, при дистанции более 6 вероятность атаки превосходит 50%. Следовательно, подход с взаимодействием с одним узлом не является безопасным и не должен использоваться на практике.

4.4. Способ связи N к N между узлами соседних реестров. Данный способ является промежуточным и обобщенным вариантом, граничные варианты которого были рассмотрены ранее. Рассмотрим процесс поиска узлов реестра-инициатора при верификации внешней транзакции.

Каждый узел реестра-акцептора поддерживает связь с K узлами родительского блокчейна. При поиске следующего реестра у каждого узла запрашивается информация о смежных реестрах – связей аналогичным образом K. В наихудшем случае атакующий возвращает

информацию только об атакующих узлах в следующем реестре $(q=1)$. Если узел честный, то вероятность того, что он вернет атакующий узел:

$$p = \frac{N_A}{N}, \tag{3}$$

где N_A – количество атакующих в следующем реестре, N – количество узлов в этом реестре. Обозначим максимальное значение вероятности:

$$\max\{p_i\} = \phi, \tag{4}$$

где p_i – вероятность (3) для реестра с номером i.

Рассмотрим цепочку реестров от текущего до целевого и математическое ожидание количества атакующих, с которыми взаимодействует один узел на каждом шаге (M). Пусть общее число узлов – N. Математическое ожидание количества атакующих в первом реестре определяется долей атакующих узлов в реестре (N_A) и числом участвующих в поиске узлов (K):

$$E\left[M^1\right] = \frac{N_A^1}{N^1} * K. \tag{5}$$

Тогда количество выбранных атакующих во втором реестре:

$$\begin{aligned} E\left[M^2\right] &= \frac{E\left[M^1\right] * K + \frac{N_A^2}{N^2} * K * \left(K - E\left[M^1\right]\right)}{K} = \\ &= E\left[M^1\right] + \left(K - E\left[M^1\right]\right) * \frac{N_A^2}{N}. \end{aligned} \tag{6}$$

В общем случае:

$$\begin{aligned} E\left[M^i\right] &= \frac{E\left[M^{i-1}\right] * K + \frac{N_A^i}{N^i} * K * \left(K - E\left[M^{i-1}\right]\right)}{K} = \\ &= E\left[M^{i-1}\right] + \frac{N_A^i}{N^i} * \left(K - E\left[M^{i-1}\right]\right). \end{aligned} \tag{7}$$

Математические ожидания образуют последовательность, причем каждый член последовательности больше, чем предыдущий

(т. к. $K > E[M]$). Разность между последовательными членами последовательности:

$$E\left[M^{i}\right] - E\left[M^{i-1}\right] = \frac{N_A^i}{N^i} * \left(K - E\left[M^{i-1}\right]\right) > 0. \quad (8)$$

Поскольку каждый элемент последовательности больше предыдущего:

$$\lim_{i\to\infty}\left(\frac{N_A^i}{N^i} * \left(K - E\left[M^{i-1}\right]\right)\right) = 0. \quad (9)$$

Поэтому выполняется условие:

$$\forall \varepsilon > 0 \exists N(\varepsilon), \forall n, m > N(\varepsilon): \left|E\left[M^{n}\right] - E\left[M^{m}\right]\right| < \varepsilon. \quad (10)$$

Следовательно, последовательность сходится по критерию Коши. Рассмотрим предел последовательности в наихудшем случае – когда вероятность атаки на каждом шаге равна максимальной вероятности атаки (ϕ). Воспользуемся методом итераций.

$$E\left[M^{i+1}\right] = \phi\left(K - E\left[M^{i}\right]\right) = \phi * K - \phi * E\left[M^{i}\right] = \varphi\left(E\left[M^{i}\right]\right) \quad (11)$$

$$\begin{gathered}\lim_{i\to\infty} E\left[M^{i+1}\right] = \lim_{i\to\infty} \varphi\left(E\left[M^{i}\right]\right) = \varphi\left(\lim_{i\to\infty} E\left[M^{i}\right]\right) \Rightarrow \\ \Rightarrow \mathrm{M} = \varphi(M) = \phi * K - \phi * M \Rightarrow M = K.\end{gathered} \quad (12)$$

Следовательно, с увеличением количества шагов до целевого реестра доля атакующих в выборке стремится к 1, и использование этого протокола на практике недопустимо. Данная ситуация подтверждает возможность атаки, описанной при рассмотрении полного графа с родительским реестром, однако в данном случае защита на основе пересечения множеств неприменима, поскольку честные узлы в общем случае возвращают разный набор узлов для следующего шага, а подконтрольные атакующему узлы – один и тот же.

5. Стойкий протокол SVP

5.1. Описание протокола. Основным результатом работы является безопасный протокол поиска и верификации внешних

транзакций, основанный на свойствах блокчейна, реализующего устойчивый распределенный реестр:

1. Свойство общего префикса (CPP): для любых двух честных узлов цепочки блоков имеют общий префикс, получаемый отсечением k блоков.
2. Свойство качества цепочки (CQP): для последовательности из l блоков доля блоков, созданных атакующими, не превосходит μ.

При этом для свойства качества цепочки выполняются важные вспомогательные соотношения [10]. При этом доля блоков, созданных атакующим в худшем случае строго меньше 1:

$$\mu = \left(1+\frac{\delta}{2}\right)*\frac{t}{n-t} < 1-\frac{\delta}{2} < 1, \tag{13}$$

где t – число атакующих, n – число узлов, δ – преимущество честных узлов. В случае протоколов, реализующих доказательство доли владения [8-9,17] величина μ определяется как:

$$\epsilon \in (0,1), \beta \in [0,1], f \in (0,1] \Rightarrow \mu = \frac{\epsilon \beta f}{16} < 1, \tag{14}$$

где ϵ - качество концентрации случайных величин в «типичных исполнениях» (понятие установлено в [10]), β – ожидаемое число найденных атакующими решений задачи доказательства работы за раунд, f – общее ожидаемое число найденных решений за раунд.

Функционал многомерного блокчейна подписывается на обновления состояния в соседнем блокчейне. При этом отслеживаются последние $l+k$ блоков. Кроме того, по мере погружения блоков на безопасную глубину функционал устанавливает взаимодействие с узлами, создавшими последние l блоков, погрузившихся на достаточно безопасную глубину. Для упрощения рассмотрения и анализа безопасности не учитывается, что l и k отличаются для разных реестров – в реальности они выбираются на каждой итерации алгоритма. При запросе на верификацию транзакции осуществляется следующая последовательность действий:

1. Генерируется $l*l$ целых случайных чисел в диапазоне (0, N), где N – индекс блока на глубине k.
2. У каждого из l узлов, с которым установлено соединение, запрашивается l блоков и последовательность хэш-сумм от каждого блока до блока, созданного узлом в пределах l блоков.

3. Для каждого блока проверяется последовательность хэш-сумм. Если хотя бы для одного из блоков хэш-сумма некорректна, отбросить все результаты от узла-источника.
4. Из всех результатов выбрать l узлов. Для каждого блока запросить у узла-источника адрес узла, создавшего блок. Запросить у узла подтверждение создания блока – зависит от конкретного протокола блокчейна и структуры блоков. Если проверка осуществлена некорректно, заменить узел на случайно выбранный из полученного на шаге 3 множества.
5. Если полученные узлы не принадлежат целевому блокчейну, то перейти к шагу 1.
6. Иначе запросить у полученных l узлов верификацию транзакции и принять решение по принципу большинства.

Стоит отметить, что на практике оповещения о новых состояниях приходят с задержкой. Поэтому допустимо использовать окна большего размера – $k + \Delta k$, $l + \Delta l$.

5.2. Доказательство безопасности протокола. Утверждение 3. Протокол поиска и верификации позволяет корректно верифицировать внешнюю транзакцию с вероятностью, близкой к 1, при правильном выборе множества опрашиваемых узлов.

По определению многомерного блокчейна все реестры в его пределах являются устойчивыми. Реестр является устойчивым в том случае, если он выполняет требования к росту цепочки (CGP), общему префиксу (CPP) и чистоте цепочки (CQP) [10]. Параметром предиката CQP являются l – подпоследовательность блоков и μ – доля блоков, созданных атакующим в подпоследовательности. Параметром предиката CPP является k – глубина, на которой для всех честных узлов имеется общий префикс.

Алгоритм отслеживает блоки, погрузившиеся глубже, чем k. Следовательно, по свойству общего префикса у всех честных узлов цепочка длиной l совпадает. Кроме того, по свойству чистоты цепочки среди всех этих блоков есть как минимум один созданный честным узлом. Рассмотрим возможные атаки на протокол со стороны атакующего.

Событие 1 – Атакующий формирует блоки при запросе информации. Вероятность того, что произвольный блок был создан атакующим, совпадает с долей атакующих систему узлов (меньше 0.5). При попытке подделки цепочки заголовков от выбранного до текущего блока атакующий должен быстро построить новую цепочку, в которой именно он контролирует конкретный блок, что невозможно с вероятностью, близкой к 1 для блоков, находящихся достаточно

глубоко в цепочке блоков. Кроме того, эту задачу атакующий должен решить по несколько раз для получения большинства в множестве мощностью l^2.

Событие 2 – Атакующий не предоставляет информацию по запросу. Это событие игнорируется. Тогда окончательное множество имеет меньшую мощность, но доля атакующих в нем не изменяется.

Следовательно, атакующий не имеет возможности повлиять на соотношение блоков, созданных честными и атакующими узлами в выбранном множестве. Поскольку из полученного множества выбираются произвольные l узлов, доля атакующих по закону больших чисел будет приблизительно равна вероятности выбора атакующих, то есть строго менее 50% за счет принципа большинства честных узлов.

Аналогичным образом при завершении поиска информация о верифицируемой транзакции запрашивается у множества из l узлов, для которых выполняется принцип большинства честных узлов.

Вероятность атаки на протокол формируется из вероятности атаки на поиск и вероятности атаки на верификацию. Успешная атака на поиск возможна только в том случае, если все узлы в выбранном на некотором шаге множестве являются атакующими:

$$\phi = \max_{\mathrm{i}} \left\{ \frac{N_A^i}{N_i} \right\}, P_i^S = \phi^l, P_A^S = 1 - \left(1 - P_i^S\right)^d = 1 - \left(1 - \phi^l\right)^d, \quad (15)$$

где P_A^S – вероятность атаки на поиск, P_i^S – вероятность атаки на процедуру поиска в реестре с индексом i, d – расстояние до целевого реестра.

Вероятность атаки на верификацию аналогичным образом определяется соотношением честных и атакующих узлов:

$$\phi = \max_{\mathrm{i}} \left\{ \frac{N_A^i}{N_i} \right\}, P_A^V = 1 - \sum_{i=0}^{\left\lceil \frac{l}{2} \right\rceil - 1} C_l^i * \phi^i * \left(1 - \phi\right)^{l-i}, \quad (16)$$

где P_A^V - вероятность атаки на верификацию. Общая вероятность атаки:

$$P_A = P_A^V + P_A^S - P_A^V * P_A^S. \quad (17)$$

Примеры значений для данной вероятности приведены в таблице 2. Следовательно, длина пути верификации не оказывает значительного влияния на вероятность компрометации. При этом

основу вероятности компрометации составляет именно вероятность компрометации верификации, а не поиска, что соответствует результатам, полученным ранее при анализе вероятности успешной верификации для взаимодействия со всеми узлами целевого реестра.

Таблица 2. Вероятность атаки

l	d	$\phi_1 = 0,5$	$\phi_2 = 0,4$	$\phi_3 = 0,3$	$\phi_4 = 0,2$	$\phi_5 = 0,1$
10	5	0,8290	0,6179	0,3504	0,1209	0,0128
10	10	0,8298	0,6181	0,3504	0,1209	0,0128
15	5	0,8491	0,5968	0,2784	0,0611	0,0022
15	10	0,8492	0,5968	0,2784	0,0611	0,0022
20	5	0,7483	0,4044	0,1133	0,0100	ε
20	10	0,7483	0,4044	0,1133	0,0100	ε

Оценка максимальных возможностей атакующего должна осуществляться для каждого используемого на практике устойчивого распределенного реестра в отдельности.

Утверждение 4. Протокол поиска и верификации транзакций, основанный на свойстве чистоты цепочки, реализует идеальный функционал поиска и верификации с вероятностью соблюдения чистоты цепочки блокчейнами, входящими в многомерный блокчейн.

В предыдущих работах была представлена модель идеального функционала для протокола поиска и верификации блоков и транзакций [6]. Этот идеальный функционал получает оповещения о новых внешних транзакциях и отвечает с задержкой, устанавливаемой атакующим. При этом вероятность корректной верификации определяется долей узлов, участвующих в верификации. Покажем, что протокол, реализующий алгоритм, представленный в данной работе, GUC-реализует указанный идеальный функционал. Для этого кратко рассмотрим последовательные гибридные модели. Исходная и целевая модели приведены на рисунке 4.

HYB0 – исходная модель, внешние взаимодействия осуществляются с использованием идеального функционала для проверки внешних транзакций.

HYB1 – модель, в которой узлы самостоятельно обеспечивают работу с поиском реестров для взаимодействия. При этом все реестры по-прежнему оповещают идеальный функционал о внешних транзакциях. В результате каждый узел с использованием протокола поиска может гарантированно обнаружить подмножество узлов реестра-инициатора, то есть поиск осуществляется самостоятельно, тогда как верификация по-прежнему осуществляется с использованием

идеального функционала. Структурно модель не изменяется, а потому для внешнего наблюдателя эквивалентна HYB0.

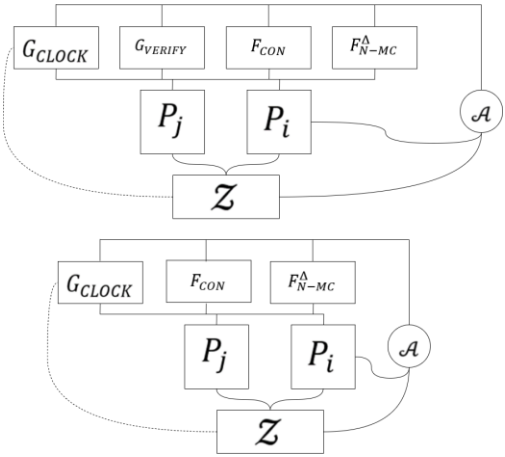


Рис. 4. GUC-модель: а) основанная на идеальном функционале; б) эквивалентная ей без данного функционала

HYB2 – разделение логики проверки транзакций. Вместо идеального функционала используется обертка, которая исполняет внутри себя набор идеальных функционалов, каждому из которых передаются запросы, связанные только с одним реестром. Внешние интерфейсы не изменяются, поэтому модель эквивалентна HYB1.

HYB3 – устранение обертки. Реестры самостоятельно взаимодействуют с идеальными функционалами. Запрос информации о том, у какого идеального функционала запрашивать верификацию, осуществляется у узлов, найденных через протокол поиска. Поскольку протокол поиска гарантированно находит узлы, поддерживающие реестр-инициатор, информация об идеальном функционале выявляется корректно с вероятностью, зависящей от доли честных узлов – Соотношение 6. При корректном подборе параметров эта вероятность стремится к 1. Стоит отметить, что вероятность некорректной верификации была учтена при построении GUC-модели для идеального функционала поиска и верификации транзакций (параметр γ).

HYB4 – запрос информации непосредственно у узлов. Аналогично HYB3 информация запрашивается у узлов, однако запрашивается именно информация о корректности транзакции. Вероятность корректной верификации при этом остается прежней,

поскольку честные узлы следуют протоколу и корректно верифицируют транзакцию. Эта модель эквивалентна протоколу поиска и верификации транзакций.

График для различных мощностей множества приведен на рисунке 5. За основу взята сеть, в которой присутствует 30 узлов. В качестве размера множества используется параметр l, введенный ранее. Следовательно, при незначительном оценочном количестве атакующих допустимо использовать как процедуру поиска, так и процедуру верификации предложенного протокола SVP. Однако, в случае если количество атакующих может достигать 50%, рекомендуется применять для процедуры верификации иные протоколы, то есть использовать комбинацию подходов.

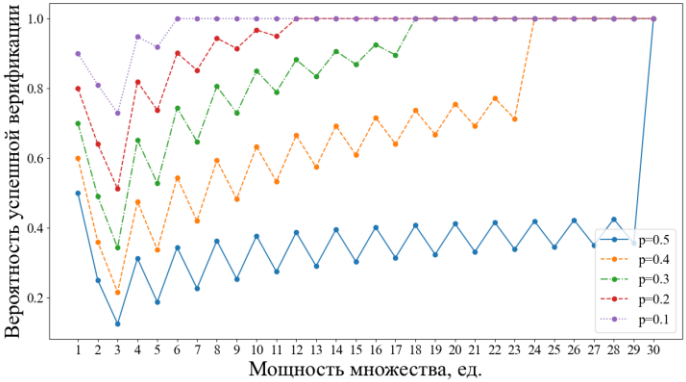


Рис. 5. Вероятность успешной верификации при различных размерах используемого подмножества

6. Комбинация подходов. Существует несколько возможных комбинаций подходов для различных условий работы многомерного блокчейна:

1. SVP и централизованный подход – в каждом реестре создается доверенный узел, отвечающий за верификацию внешних транзакций. В этом случае безопасный поиск осуществляется с помощью SVP, а верификация с использованием этого узла.
2. SVP и полный граф – если доля атакующих может достигать 50%, то допустимо организовывать поиск узлов с использованием протокола поиска и верификации, а затем – увеличить мощность опрашиваемого множества для повышения вероятности успешной верификации.

Стоит отметить, что полученные результаты не противоречат полученным ранее в иных работах, посвященных обмену информацией между блокчейнами [11-16]. Этап поиска завершается с вероятностью, близкой к 1. Этап верификации может быть завершен с вероятностью, близкой к 1 в случае, если количество атакующих сравнительно невелико (до 20%). Данные результаты схожи с результатами, достигнутыми в механизме достижения консенсуса Ripple. При использовании решений [5] или [11] возможно подтверждение корректности транзакций с вероятностью, близкой к 1, и за сравнительно небольшое время.

7. Обсуждение результатов. В работе было рассмотрено несколько подходов к построению протокола поиска и верификации блоков и транзакций. Основным результатом работы является протокол SVP, позволяющий осуществлять безопасный обмен данными между самостоятельными устойчивыми распределенными реестрами. Для всех рассмотренных подходов к построению протокола поиска и верификации проведен анализ безопасности. Рассмотрим основные отличия предлагаемого решения от существующих аналогов.

В работе [6] была рассмотрена модель многомерного блокчейна, использующая центральный модуль для верификации транзакций. Основным отличием предложенного подхода в части верификации является отсутствие требования к атомарности. Все внешние транзакции осуществляются в две фазы. Преимуществом данного подхода является независимость реестров друг от друга при принятии транзакций: реестр-акцептор принимает транзакцию в подходящий для поддерживающих его узлов момент времени.

Предложенные ранее в литературе подходы к обмену данными между устойчивыми распределенными реестрами [7-12] были предназначены для проведения платежей между криптовалютами, построенными на основе технологии блокчейн. Существенным отличием протокола поиска и верификации является независимость от приложения. Поэтому он может быть использован не только в сфере криптовалют, но и в любых приложениях, построенных с использованием технологии распределенного реестра.

Другое отличие от принципов, предложенных авторами работ, посвященных сайдчейнам, заключается в принципе отслеживания цепочки блоков соседнего блокчейна. В случае сайдчейнов создаются специальные контрольные точки, которые позволяют быстрее находить требуемые блоки, содержащие транзакции, и верифицировать их. В случае протокола поиска и верификации блоков и транзакций узлы

находятся в постоянном взаимодействии с узлами соседнего реестра и отслеживают цепочку из $2k$ последних блоков.

Отдельно стоит отметить, что предъявляемые к решению требования фактически совпадают с аналогичными требованиями, выдвинутыми в [5]. Решение удовлетворяет основным требованиям, предъявляемым к сайдчейнам и протоколам обмена информации между ними.

Ключевым преимуществом решения в сравнении с сайдчейнами можно считать возможность поиска узлов, поддерживающих реестр-инициатор. В случае сайдчейнов информация об узлах соседнего реестра подается протоколу на вход. В случае многомерного блокчейна (и протокола поиска и верификации блоков и транзакций) реестры находятся в иерархической зависимости, а потому могут находить узлы, поддерживающие произвольный реестр, по мере необходимости. Также благодаря использованию технологии в процессе проверки внешних транзакций может участвовать больше двух реестров.

Наиболее существенным достоинством предложенного решения является отсутствие необходимости в модификации реализации существующих решений (за исключением поддержки внешних транзакций и процедуры верификации). Подход не зависит от особенностей работы конкретных систем и позволяет им функционировать самостоятельно в пределах многомерного блокчейна. Существенной особенностью решения является отсутствие предикатов в отличие от рассмотренных существующих аналогов: у узлов сайдчейна запрашивается информация о корректности транзакции, а не доказательство корректности. Поэтому возможна работа с произвольными механизмами достижения консенсуса [22-24].

Стоит отметить, что при количестве атакующих, близком к 50% вероятность компрометации системы достаточно велика при опросе подмножества узлов. Поэтому в случаях, когда допустимое количество компрометаций не определено, процедура верификации должна быть построена на основе [5] или [11]. Иначе допустимо применение более простого подхода, рассмотренного в данной работе.

8. Заключение. В работе рассмотрена проблема безопасного обмена данными между самостоятельными устойчивыми распределенными реестрами в пределах многомерного блокчейна. Рассмотрены различные подходы к построению протокола поиска и верификации внешних транзакций. Для этих подходов произведена оценка безопасности. Предложен подход к построению безопасного протокола поиска и верификации транзакций, доказана его

безопасность и эквивалентность идеальному функционалу поиска и верификации транзакций.

Целью дальнейшей работы является построение новых механизмов для проведения внешних транзакций, оценка безопасности таких технологий, как кэширование, при их использовании в протоколе поиска и верификации блоков и транзакций. Кроме того, перспективным направлением для исследований является внедрение криптографии с нулевым разглашением в многомерный блокчейн и адаптация протокола поиска и верификации блоков и транзакций для проверки транзакций с нулевым разглашением. Также возможна адаптация криптографических алгоритмов с нулевым разглашением для проведения внешних транзакций.

Наконец, важным направлением для практических исследований является сопоставление многомерного блокчейна с другими технологиями с точки зрения различных характеристик работы. По результатам исследований должна быть произведена оптимизация характеристик и усовершенствование существующих протоколов для улучшения характеристик его работы.

Литература

1. *Шилов И.М., Заколдаев Д.А.* Многомерный блокчейн и его преимущества // Информационные технологии. 2020. Т. 26. № 6. С. 360–367.
2. *Badertscher C., Maurer U., Tschudi D., Zikas V.* Bitcoin as a Transaction Ledger: A Composable Treatment // Advances in Cryptology – CRYPTO 2017. 2017. pp. 324-356.
3. *Vukolic M.* Rethinking permissioned blockchains // Proceedings of the ACM Workshop on Blockchain, Cryptocurrencies and Contracts. 2017. pp. 3-7.
4. *Cachin C., Guerraoui R., Rodrigues L.* Introduction to Reliable and Secure Distributed Programming // Springer-Verlag, Berlin, Heidelberg. 2011. P. 279.
5. *Pease M., Shostak R., Lamport L.* Reaching agreement in the presence of faults // Journal of the ACM. 1980. vol. 27. pp. 228-234.
6. *Шилов И.М., Заколдаев Д.А.* Модель устойчивого распределенного реестра для анализа безопасности многомерного блокчейна // Научно-технический вестник информационных технологий, механики и оптики. 2021. Т. 21. №2. С. 249-255.
7. *Garay J., Kiayias A., Leonardos N.* The Bitcoin Backbone Protocol: Analysis and Applications // Advances in Cryptology - EUROCRYPT 2015. 2015. vol. 9057. pp. 281-310.
8. *Badertscher C., Gaži P., Kiayias A., Russell A., Zikas V.* Ouroboros Genesis: Composable Proof-of-Stake Blockchains with Dynamic Availability // ACM Conference on Computer and Communications Security – ACM CCS 2018. 2018. pp. 913–930.
9. *David B., Gaži P., Kiayias A., Russell A.* Ouroboros Praos: An Adaptively-Secure, Semi-synchronous Proof-of-Stake Blockchain // Advances in Cryptology – EUROCRYPT 2018. 2018. vol. 10821. pp. 66-98.
10. *Garay J., Kiayias A., Leonardos N.* The Bitcoin Backbone Protocol with Chains of Variable Difficulty // Advances in Cryptology – CRYPTO 2017. 2017. vol. 10401. pp. 291-323.

11. *Kiayias A., Lamprou N., Stouka AP.* Proofs of Proofs of Work with Sublinear Complexity // Financial Cryptography and Data Security. 2016. vol. 9604. pp. 61-78.
12. *Kiayias A., Miller A., Zindros D.* Non-interactive Proofs of Proof-of-Work // Financial Cryptography and Data Security. 2020. vol. 12059. pp. 505-522.
13. *Back A., Corallo M., Dashjr L., Friedenbach M., Maxwell G., Miller A., Poelstra A., Timon J., Wuille P.* Enabling Blockchain Innovations with Pegged Sidechains. URL: https://blockstream.com/sidechains.pdf (дата обращения: 29.04.2021).
14. *Gazi P., Kiayias A., Zindros D.* Proof-of-Stake Sidechains // 2019 IEEEE Symposium on Security and Privacy (SP). 2019. vol. 1. pp. 677-694.
15. *Sompolinsky Y., Zohar A.* Accelerating Bitcoin's Transaction Processing Fast Money Grows on Trees, Not Chains // IACR Cryptology ePrint Archive. 2013.
16. *Singh A., Click K., Parizi R.M., Zhang Q., Dehghantanha A., Choo K.K.R.* Sidechain technologies in blockchain networks: An examination and state-of-the-art review // Journal of Network and Computer Applications. 2020. vol. 149.
17. *Kiayias A., Russell A., David B., Oliynykov R.* Ouroboros: A Provably Secure Proof-of-Stake Blockchain Protocol // Advances in Cryptology – CRYPTO 2017. 2017. vol. 10401. pp. 357-388.
18. *Canetti R.* Universally composable security: a new paradigm for cryptographic protocols // Proceedings 42nd IEEE Symposium on Foundations of Computer Science. 2001. pp. 136-145.
19. *Canetti R.* Universally composable signatures, certification, and authentication // Proceedings of 17th Computer Security Foundations Workshop (CSFW). 2014. pp. 219-235.
20. *Canetti R., Dodis Y., Pass R., Walfish S.* Universally Composable Security with Global Setup // Theory of Cryptography. 2007. vol. 4392. pp. 61-85.
21. *Canetti R., Shahaf D., Vald M.* Universally Composable Authentication and Key-Exchange with Global PKI // Public-Key Cryptography – PKC 2016. 2016. vol. 9615. pp. 265-296.
22. *Bentov I., Gabizon A., Mizrahi A.* Cryptocurrencies Without Proof of Work // Financial Cryptography and Data Security. 2016. vol. 9604. pp. 142-157.
23. *David B., Dowsley R., Larangeira M.* ROYALE: A Framework for Universally Composable Card Games with Financial Rewards and Penalties Enforcement // Financial Cryptography and Data Security. vol. 11598. pp. 282-300.
24. *Duan S., Meling H., Peisert S., Zhang H.* BChain: Byzantine Replication with Hight Throughput and Embedded Reconfiguration // Principles of Distributed Systems – OPODIS 2014. 2014. vol. 8878. pp. 91-106.

Шилов Илья Михайлович — научный сотрудник, Университет ИТМО. Область научных интересов: устойчивые распределенные реестры, многомерный блокчейн, анализ данных и математическая статистика в информационной безопасности. Число научных публикаций – 13. ilia.shilov@yandex.ru; Кронверкский пр., д. 49, г. Санкт-Петербург, 197101, РФ; р.т. +7(931)3604143.

Заколдаев Данил Анатольевич — к.т.н., доцент, декан факультета безопасности информационных технологий, Университет ИТМО. Область научных интересов: САПР, безопасность цифрового производства, аппаратные средства защиты информации, информационные технологии в образовании. Число научных публикаций – 200. d.zakoldaev@itmo.ru; Кронверкский пр., д. 49, г. Санкт-Петербург, 197101, РФ.

Поддержка исследований. Работы выполнены при поддержке ФГБУ «Фонд содействия развитию малых форм предприятий в научно-технической сфере» (договор № 14492ГУ/2019 от 18.07.2019).

DOI 10.15622/ia.20.4.2

I. SHILOV, D. ZAKOLDAEV

SECURITY OF SEARCH AND VERIFICATION PROTOCOL IN MULTIDIMENSIONAL BLOCKCHAIN

Shilov I., Zakoldaev D. **Security of Search and Verification Protocol in Multidimensional Blockchain.**

Abstract. The issue of secure data exchange and performing external transactions between robust distributed ledgers has recently been among the most significant in the sphere of designing and implementing decentralized technologies. Several approaches have been proposed to speed up the process of verifying transactions on adjacent blockchains. The problem of search has not been under research yet. The paper contains security evaluation of data exchange between independent robust distributed ledgers inside multidimensional blockchain. Main principles, basic steps of the protocol and major requirements for it are observed: centralized approach, subset principle and robust SVP. An equivalence of centralized approach and ideal search and verification functionality is proven. The probability of successful verification in case of using fully connected network graph or equivalent approach with fully connected graph between parent and child blockchain is shown. The insecurity of approach with one-to-one links between child and parent ledgers or with a subset principle is proven. A robust search and verification protocol for blocks and transactions based on the features of robust distributed ledgers is presented. The probability of attack on this protocol is mostly defined by the probability of attack on verification and not on search. An approach to protection against an attacker with 50% of nodes in the network is given. It is based on combination of various search and verification techniques.

Keywords: Search and Verification Protocol, Blockchain, Sidechain, Multidimensional Blockchain, GUC-Framework, Robust Distributed Ledger.

Shilov Ilya – Researcher, ITMO University. Research interests: robust distributed ledgers, multidimensional blockchain, data analysis and statistics in information security. The number of publications – 13. ilia.shilov@yandex.ru; 49, Kronverksky pr., St. Petersburg, 197101, Russia; office phone: +7(931)3604143.

Zakoldaev Danil – Ph.D., Dr. Sci., Associate professor, Dean of Faculty of Secure Information Technologies, ITMO University. Research interests: CAD, cybersecurity, information security hardware, information technologies in education. The number of publications – 200. d.zakoldaev@itmo.ru; 49, Kronverksky pr., St. Petersburg, 197101, Russia.

Acknowledgements. The research is supported by Foundation for Assistance to Small Innovative Enterprises (FASIE) (contract No. 14492ГУ/2019, 18.07.2019).

References

1. Shilov I.M., Zakoldaev D.A. [Multidimensional blockchain and its advantages]. *Informacionnye tehnologii*. 2020. no. 6. pp. 360-367.
2. Badertscher C., Maurer U., Tschudi D., Zikas V. Bitcoin as a Transaction Ledger: A Composable Treatment. Advances in Cryptology – CRYPTO 2017. 2017. pp. 324-356.
3. Vukolic M. Rethinking permissioned blockchains. Proceedings of the ACM Workshop on Blockchain, Cryptocurrencies and Contracts. 2017. pp. 3-7.
4. Cachin C., Guerraoui R., Rodrigues L. Introduction to Reliable and Secure Distributed Programming. *Springer-Verlag, Berlin, Heidelberg*. 2011. p. 279.

5. Pease M., Shostak R., Lamport L. Reaching agreement in the presence of faults. *Journal of the ACM*. 1980. vol. 27. pp. 228-234.
6. Shilov I.M., Zakoldaev D.A. [The robust distributed ledger model for a multidimensional blockchain security analysis]. *Nauchno-tehnicheskij vestnik informacionnyh tehnologij, mehaniki i optiki*. 2021. vol. 21. no. 2. pp. 249-255.
7. Kiayias A., Lamprou N., Stouka AP. Proofs of Proofs of Work with Sublinear Complexity. *Financial Cryptography and Data Security*. 2016. vol. 9604. pp. 61-78.
8. Kiayias A., Miller A., Zindros D. Non-interactive Proofs of Proof-of-Work. *Financial Cryptography and Data Security*. 2020. vol. 12059. pp. 505-522.
9. Back A., Corallo M., Dashjr L., Friedenbach M., Maxwell G., Miller A., Poelstra A., Timon J., Wuille P. Enabling Blockchain Innovations with Pegged Sidechains. URL: https://blockstream.com/sidechains.pdf (дата обращения: 29.04.2021).
10. Gazi P., Kiayias A., Zindros D. Proof-of-Stake Sidechains. 2019 IEEE Symposium on Security and Privacy (SP). 2019. vol. 1. pp. 677-694.
11. Sompolinsky Y., Zohar A. Accelerating Bitcoin's Transaction Processing Fast Money Grows on Trees, Not Chains. IACR Cryptology ePrint Archive. 2013.
12. Singh A., Click K., Parizi R.M., Zhang Q., Dehghantanha A., Choo K.K.R. Sidechain technologies in blockchain networks: An examination and state-of-the-art review. *Journal of Network and Computer Applications*. 2020. vol. 149.
13. Kiayias A., Russell A., David B., Oliynykov R. Ouroboros: A Provably Secure Proof-of-Stake Blockchain Protocol. Advances in Cryptology – CRYPTO 2017. 2017. vol. 10401. pp. 357-388.
14. Canetti R. Universally composable security: a new paradigm for cryptographic protocols. Proceedings 42nd IEEE Symposium on Foundations of Computer Science. 2001. pp. 136-145.
15. Canetti R. Universally composable signatures, certification, and authentication. Proceedings of 17th Computer Security Foundations Workshop (CSFW). 2014. pp. 219-235.
16. Canetti R., Dodis Y., Pass R., Walfish S. Universally Composable Security with Global Setup. Theory of Cryptography. 2007. vol. 4392. pp. 61-85.
17. Canetti R., Shahaf D., Vald M. Universally Composable Authentication and Key-Exchange with Global PKI. Public-Key Cryptography – PKC 2016. 2016. vol. 9615. pp. 265-296.
18. Garay J., Kiayias A., Leonardos N. The Bitcoin Backbone Protocol: Analysis and Applications. Advances in Cryptology - EUROCRYPT 2015. 2015. vol. 9057. pp. 281-310.
19. Badertscher C., Gaži P., Kiayias A., Russell A., Zikas V. Ouroboros Genesis: Composable Proof-of-Stake Blockchains with Dynamic Availability. ACM Conference on Computer and Communications Security – ACM CCS 2018. 2018. pp. 913–930.
20. David B., Gaži P., Kiayias A., Russell A. Ouroboros Praos: An Adaptively-Secure, Semi-synchronous Proof-of-Stake Blockchain. Advances in Cryptology – EUROCRYPT 2018. 2018. vol. 10821. pp. 66-98.
21. Garay J., Kiayias A., Leonardos N. The Bitcoin Backbone Protocol with Chains of Variable Difficulty. Advances in Cryptology – CRYPTO 2017. 2017. vol. 10401. pp. 291-323.
22. Bentov I., Gabizon A., Mizrahi A. Cryptocurrencies without Proof of Work. *Financial Cryptography and Data Security*. 2016. vol. 9604. pp. 142-157.
23. David B., Dowsley R., Larangeira M. ROYALE: A Framework for Universally Composable Card Games with Financial Rewards and Penalties Enforcement. *Financial Cryptography and Data Security*. vol. 11598. pp. 282-300.
24. Duan S., Meling H., Peisert S., Zhang H. BChain: Byzantine Replication with Hight Throughput and Embedded Reconfiguration. *Principles of Distributed Systems* – OPODIS 2014. 2014. vol. 8878. pp. 91-106.

УДК 004.056 **DOI** 10.15622/ia.20.4.3

Д.П. Зегжда, М.О. Калинин, В.М. Крундышев, Д.С. Лаврова,
Д.А. Москвин, Е.Ю. Павленко

ПРИМЕНЕНИЕ АЛГОРИТМОВ БИОИНФОРМАТИКИ ДЛЯ ОБНАРУЖЕНИЯ МУТИРУЮЩИХ КИБЕРАТАК

Зегжда Д.П., Калинин М.О., Крундышев В.М., Лаврова Д.С., Москвин Д.А., Павленко Е.Ю. **Применение алгоритмов биоинформатики для обнаружения мутирующих кибератак.**

Аннотация. Функционал любой системы может быть представлен в виде совокупности команд, которые приводят к изменению состояния системы. Задача обнаружения атаки для сигнатурных систем обнаружения вторжений эквивалентна сопоставлению последовательностей команд, выполняемых защищаемой системой, с известными сигнатурами атак. Различные мутации в векторах атак (включая замену команд на равносильные, перестановку команд и их блоков, добавление мусорных и пустых команд) снижают эффективность и точность обнаружения вторжений. В статье проанализированы существующие решения в области биоинформатики, рассмотрена их применимость для идентификации мутирующих атак. Предложен новый подход к обнаружению атак на основе технологии суффиксных деревьев, используемой при сборке и проверке схожести геномных последовательностей. Применение алгоритмов биоинформатики позволяет добиться высокой точности обнаружения мутирующих атак на уровне современных систем обнаружения вторжений (более 90%), при этом превосходя их по экономичности использования памяти, быстродействию и устойчивости к изменениям векторов атак. Для улучшения показателей точности проведен ряд модификаций разработанного решения, вследствие которых точность обнаружения атак увеличена до 95% при уровне мутаций в последовательности до 10%. Метод может применяться для обнаружения вторжений как в классических компьютерных сетях, так и в современных реконфигурируемых сетевых инфраструктурах с ограниченными ресурсами (Интернет вещей, сети киберфизических объектов, сенсорные сети).

Ключевые слова: алгоритм Укконена, безопасность, биоинформатика, выравнивание, мутация, обнаружение вторжений, полиморфизм, сигнатура, суффиксное дерево.

1. Введение. Современные системы обнаружения вторжений (СОВ) выполняют функцию пассивной защиты отдельных хостов сети (хостовые СОВ) или сетевых соединений (сетевые СОВ), ведя наблюдение и анализ событий безопасности, происходящих внутри системы на уровне системных вызовов или сетевых потоков [1]. СОВ фиксируют признаки различных видов вредоносной активности, такие как эксплуатация программных уязвимостей, DDoS-атаки, сканирование портов и попытки проникновения в сеть. Независимо от типа к СОВ предъявляются высокие требования по качеству обнаружения и скорости распознавания атак, так как от этих характеристик зависит наносимый атакой урон [2, 3].

СОВ постоянно собирают различную информацию и составляют набор идентифицирующих признаков, анализируя который делают вы-

вод о текущей безопасности контролируемой системы. Распространенные сигнатурные СОВ для выявления вторжений сравнивают текущее состояние системы с известными шаблонами (сигнатурами) небезопасных состояний. Если текущее состояние совпадает с одной из сигнатур, СОВ сигнализируют о выявленном вторжении. В отличие от СОВ, построенных на основе оценки аномалий, в которых определяют отклонения от среднестатистического профиля поведения контролируемой системы, сигнатурные СОВ имеют высокие показатели качества и скорости обнаружения, низкий уровень ошибок первого и второго рода и не требуют выделенного этапа профилирования [4-6].

Определим защищаемую систему как набор сущностей, которые взаимодействуют друг с другом. Информационное взаимодействие между сущностями в системе реализуется путем выполнения команд. В результате взаимодействия изменяется состояние системы. Знание о том, что в интервале времени были в некоторой последовательности выполнены команды, определяет все состояния системы, получаемые из исходного. Поведение, которое представлено в виде последовательности команд, приводящих защищаемую систему в небезопасное состояние, будем называть атакой. Задача обнаружения вторжений для сигнатурной СОВ эквивалентна сопоставлению текущих последовательностей команд с известными сигнатурами атак, представленными последовательностями команд, ведущих к небезопасному состоянию. При сопоставлении последовательностей проблему, характерную для сигнатурного подхода, представляет наличие мутаций, или полиморфизм, атак [7-9]. Мутации усложняют обнаружение вторжений, поскольку требуют поддержания чрезмерно больших баз сигнатур для всех возможных вариаций атак, чрезмерных ресурсозатрат на поиск и сопоставление сигнатур с наблюдаемыми последовательностями, своевременного и постоянного пополнения сигнатур атак [10].

Мутации атак включают такие изменения в последовательности команд, как: замена команд на равносильные, перестановка команд и их блоков в цепочке, добавление мусорных и пустых команд. Например, мутирующие последовательности для цепочки $S = abcdefgh$: $S_1 = aBcdefgh$ (замена равносильной командой), $S_2 = abcedfgh$ (перестановка команд), $S_3 = abc\_edfgh$ (добавление мусорных команд, включая пустые команды и пропуски во времени). Поскольку для сигнатурных СОВ единственный способ определения атаки – нахождение идентичной последовательности в списке сигнатур, то даже незначительные изменения в последовательностях команд не позволяют сигнатурной СОВ распознать мутирующие атаки с использованием шаблонов.

Таким образом, сигнатурные СОВ уязвимы к полиморфизму атак и зависимы от наполнения базы сигнатур. Решением данной проблемы может служить быстродействующий механизм, способный с высокой точностью обрабатывать и сопоставлять поступающие последовательности команд с большим набором имеющихся сигнатур, несмотря на мутации в текущей последовательности. Для решения поставленной задачи в статье рассматриваются следующие аспекты:

– проанализированы существующие алгоритмы биоинформатики, которые позволяют решить аналогичную природную задачу – локализовать совпадения между последовательностью биокодов о состоянии биологической системы (генома) с одним из элементов объемной базы геномных сигнатур, а также рассмотрена их применимость для выявления мутирующих атак сигнатурными СОВ;

– предложен новый сигнатурный метод обнаружения мутирующих атак, основанный на механизме суффиксных деревьев, используемых при сборке и проверке схожести геномных последовательностей.

2. Исследование возможных решений. Требования, схожие с теми, что выдвигаются для сигнатурных СОВ относительно противодействия мутирующим атакам, предъявляются к алгоритмам биоинформатики, целью которых является сборка и сопоставление геномных последовательностей. Задачей обработки геномных последовательностей в биоинформатике является восстановление и упорядочивание больших цепочек ДНК длиной до миллиардов нуклеотидных кодов на основании информации, полученной в результате секвенирования [11-13]. Секвенирование – общее название биоинформационных методов, которые позволяют установить последовательность нуклеотидов в последовательности ДНК. В настоящее время нет ни одного метода секвенирования, который бы работал над геномными последовательностями целиком – все они устроены таким образом, что сначала готовится большое число нуклеотидных блоков (геном многократно клонируется и разрезается на блоки – риды), которые затем обрабатываются. Методы обработки ридов отличаются вариантами параллелизма и организации вычислений на структурах данных кодированного представления ридов.

Риды – последовательности, получаемые при секвенировании и содержащие информацию о фрагментах генома. Рид представляется строкой четырехбуквенного алфавита нуклеотидов, соответствующей фрагменту генома. Секвенаторы, в зависимости от механизма их работы, совершают ошибки, наиболее частая из которых – замена одного нуклеотида на другой (например, в риде находится нуклеотид Т, а в соответствующей позиции генома – А), а также ошибки пропуска ряда

нуклеотидов и вставки посторонних нуклеотидов в рид. При сборке генома из ридов необходимо устранить ошибки секвенирования. Для этого используются различные алгоритмы выравнивания последовательностей и поиска гомологически близких строк геномных кодов. Тем самым сборка и выравнивание геномов аналогичны задаче поиска подобных последовательностей среди сигнатур СОВ. Цель алгоритмов сборки генома, как и сигнатурных СОВ, состоит в нахождении совпадений рида (в СОВ – участка в последовательности команд системы) с одним из участков из объемной базы сигнатур. Алгоритмы биоинформатики полностью отвечают требованиям, предъявляемым к СОВ, устойчивым к мутирующим атакам, так как изменения, вносимые в атаки, проявляются в последовательности команд как вставка дополнительных элементов в последовательность и аналогичны ошибкам вставки лишних блоков нуклеотидов секвенатором. Качество и быстродействие биоинформационных алгоритмов обусловлено требуемой от них точностью и скоростью обработки больших последовательностей [14].

Классификация известных биоинформационных алгоритмов обработки геномных последовательностей представлена на рисунке 1.

Семейство алгоритмов De Novo [15] используется для сборки ранее неизвестного генома и основано на избыточности ридов, за счет которой восстанавливается порядок их следования. Алгоритмы Overlap Layout Concensus [16] и алгоритмы на графах де Брюина [17, 18] обладают квадратичной сложностью и поэтому не подходят для быстрой обработки больших последовательностей.

Алгоритмы выравнивания последовательностей на входе помимо набора ридов получают ранее восстановленный геном, который был собран до этого [19]. Оценивание схожести последовательностей, выполняемое при этом, упрощает процесс нахождения мутаций. Глобальное выравнивание [20] предполагает, что последовательности изначально гомологичны, и учитывает это сходство на протяжении всего выравнивания. Например, глобальное выравнивание Нидлмана-Вунша [21, 22] основано на оценке коллинеарности двух последовательностей и работает оптимально при сравнении очень схожих последовательностей. Соответствие выравненных символов задается матрицей схожести, значения ячеек которой растут при совпадении и уменьшаются при несовпадении элементов. Корректная работа алгоритма обусловлена свойством аддитивности, по которому делается оптимальный выбор на каждом шаге.

Локальное выравнивание ищет схожие блоки в последовательностях и выравнивает последовательности относительно блоков. Соответ-

ственно, для пары последовательностей может быть несколько локальных выравниваний. Алгоритм Смита-Уотермана [23, 24] учитывает локальные выравнивания схожих областей, которые попадают в разные наборы последовательностей. В отличие от схемы Нидлмана-Вунша значения в матрице схожести не могут опускаться ниже определенного минимума, и, таким образом, итоговое выравнивание разбивается на несколько оптимальных участков.

Множественное выравнивание трех и более геномных последовательностей предполагает, что входной набор последовательностей имеет эволюционную связь. Ввиду большей вычислительной сложности по сравнению с парным выравниванием, многие реализации множественного выравнивания используют эвристические алгоритмы.

Рис. 1. Классификация биоинформационных алгоритмов обработки последовательностей

Различают следующие алгоритмы, реализующие множественное выравнивание:

– прогрессивный алгоритм выравнивания [25, 26] реализует две стадии:

а) построение бинарного путеводного дерева, в котором листья являются последовательностями,

б) построение множественного выравнивания путем добавления последовательностей к растущему выравниванию согласно путеводному дереву.

Выравнивание может быть неудачным в случае набора сильно отдаленных друг от друга последовательностей. Ошибки, полученные на любой стадии растущего множественного выравнивания, доходят до результирующего выравнивания. Алгоритм требователен к схожести начальных последовательностей, что не подходит для решения поставленной задачи;

– итеративный алгоритм [25, 27] работает аналогично прогрессивному, но при этом он неоднократно перестраивает исходные выравнивания при добавлении новых последовательностей. Алгоритм может возвращаться к первоначально посчитанным парным выравниваниям и подвыравниваниям, содержащим подмножества последовательностей из запроса, и таким образом оптимизировать целевую функцию и повышать качество. Прогрессивное и итеративное выравнивания достаточно эффективны для одновременной обработки большого числа (100…1000) последовательностей, но, так как они эвристические, то они не гарантируют нахождения глобального оптимального выравнивания;

– скрытая марковская модель [12, 28] – вероятностная модель, которая может оценить правдоподобие для всех возможных комбинаций пропусков, совпадений или несовпадений для того, чтобы определить наиболее вероятное множественное выравнивание. Скрытая марковская модель может вычислять одно выравнивание с высоким весом, но также может сгенерировать семейство возможных выравниваний, которые затем могут быть оценены по их весу. Модель может быть использована для получения как глобальных, так и локальных выравниваний. Несмотря на то, что решение на базе скрытой марковской модели появилось сравнительно недавно, оно значительно улучшило вычислительную сложность, особенно для последовательностей, содержащих перекрывающиеся области.

Алгоритмы, основанные на скрытых марковских моделях, представляют множественное выравнивание в виде направленного ациклического графа, который состоит из серий узлов, представляющих собой возможные состояния в колонках выравнивания. В этом представлении абсолютно консервативная колонка (то есть последовательности во множественном выравнивании имеют в этой позиции определенный символ) кодируется как один узел со множеством исходящих соединений с символами, возможными в следующей позиции выравнивания. В терминах стандартной скрытой марковской модели наблюдаемые состояния – отдельные колонки выравнивания, а «скрытые» состояния представляют собой предполагаемую предковую последовательность, от которой последовательности из входного набора могли произойти. Аналогично прогрессивному выравниванию алгоритм требователен к

классификации сигнатур, но представляет более близкое к конкретной сигнатуре выравнивание.

Оптимизационные алгоритмы вычислительного интеллекта также используются для построения множественных выравниваний. Например, генетический алгоритм реализует гипотетическое эволюционное разделение серий возможных цепочек на фрагменты и повторную их перестройку с вводом разрывов в различные локации [29]. Алгоритм решает задачу поэтапного приближения к шаблону, что удовлетворяет цели сборки генома, но не позволяет быстро оценить схожесть пар последовательностей, что делает его непригодным для обнаружения вторжений.

Отдельно выделяют быстрые алгоритмы локального парного выравнивания – выравниватели коротких прочтений:

– алгоритм на базе хэш-таблиц, который использует хеш-функцию, трансформирующую строку в ключ быстрого поиска [30]. Наиболее простым способом было бы разбиение последовательности генома на слова, совпадающие по длине с ридом, но этот подход не работает, так как длинные слова обычно уникальны и их хранение требует слишком много места в памяти. Вместо этого используются хеширование более коротких блоков, которые встречаются гораздо чаще. После того, как с помощью хеш-функции получены подходящие позиции, можно картировать оставшуюся часть прочтения на геном. Подход разделения прочтения на несколько частей позволяет заложить в алгоритме возможность замен. Рид можно разбить на множество последовательностей со сдвигом в несколько нуклеотидов. Таким образом можно бороться с ошибками секвенирования (ошибочный фрагмент с обеих сторон окружен короткими последовательностями, которые в свою очередь успешно выравниваются).

Алгоритмы хеширования плохо справляются с повторами, так как сильно растет количество ридов, которое необходимо проверять. Для решения этой проблемы были разработаны суффиксные деревья [31, 32]. Преимущество суффиксных деревьев заключается в том, что повторы не увеличивают время работы этого алгоритма, так как повторяющиеся участки схлопываются в суффиксном дереве. Данный механизм работает крайне быстро при условии отсутствия ошибок и замен.

Сравнительная таблица 1 отражает основные свойства рассмотренных алгоритмов биоинформатики. Для дальнейшего исследования отобран механизм суффиксных деревьев, который дополнен предшествующим разбиением последовательностей на меньшие блоки для лучшей работы с мутирующими атаками. Такое решение избавляет от влияния мутационных вставок, пропусков и смены порядка команд, так как

в случае мутаций позволяет полностью восстановить из последовательности одну из наиболее близких сигнатур, находящихся в базе сигнатур.

Таблица 1 – Сравнение биоинформационных алгоритмов сборки и выравнивания геномов (m – длина анализируемого блока, n – длина генома, k – размер базы геномов, c – мощность алфавита геномов)

Алгоритм	Выравнивание мутированных последовательностей	Требовательность к чистым сигнатурам	Требовательность к большим базам сигнатур	Сложность построения структур данных	Сложность поиска	Объем памяти	Высокий коэффициент выравнивания
Алгоритм Ниддлмана-Вунша	–	+	+	$O(nmk)$	$O(nmk)$	$O(nmk)$	–
Алгоритм Смита-Ватермана	+	+	–	$O(nmk)$	$O(nmk)$	$O(nmk)$	–
Прогрессивный алгоритм	–	–	+	$O(n^2k^2)$	$O(m)$	$O(nk)$	–
Итеративный алгоритм	+	–	+	$O(n^2k^2)$	$O(m)$	$O(nk)$	+
Скрытая марковская модель	+	–	+	$O(nc^2)$	$O(m)$	$O(nk)$	+
Хэш-таблица	+	+	–	$O(nmk)$	$O(k)$	$O(nk)$	+
Суффиксное дерево	+	+	–	$O(nk)$	$O(mk)$	$O(nk)$	+

3. Метод обнаружения мутирующих атак на базе суффиксных деревьев. Бор – структура данных для хранения набора закодированных последовательностей, представляющая собой подвешенное дерево с символами на ребрах. Строки получаются последовательной записью всех символов, хранящихся на ребрах между корнем и терминальной вершиной. Размер бора линейно зависит от суммы длин всех строк. Поиск в бору занимает время, пропорциональное длине образца [33].

Рассмотрим бор, содержащий некоторый набор слов $s_1,...,s_k$. Количество вершин бора может достигать суммарной длины слов $|s_1|+...+|s_k|$. Для сокращения количества вершин рассмотрим такую цепочку вершин бора, что из каждой вершины исходит единственное ребро в следующую, и сожмем такую цепочку в одно ребро, а вместо буквы напишем на нем всю последовательность букв с ребер, которые

мы заменили. Эта последовательность букв является подстрокой некоторой строки s_i из набора, поэтому запишем на ребре только номер строки, а также начало и конец соответствующей подстроки. Сжатие всех строк в боре позволяет построить сжатый бор – корневое дерево, на каждом ребре которого написана непустая строка, обладающее следующими свойствами [33]:

– ни из какой вершины не выходят два ребра, строки для которых начинаются на один и тот же символ;

– если вершина не является корнем или листом дерева, из нее выходит не менее двух ребер.

Количество вершин в сжатом боре составляет $O(k)$, где k – количество строк в наборе. Суммарное количество вершин не превосходит $2k$. Сжатый бор занимает $O(k)$ памяти, однако для операций с ним необходимо явно хранить все строки s_i, поэтому по памяти аналогичный выигрыш не достигается.

Суффиксное дерево строки s – сжатый бор, построенный на всех суффиксах s. Такой бор занимает $O(|s|)$ памяти. Однако явное хранение всех суффиксов по отдельности не требуется: они все присутствуют в строке s. Это значит, что суффиксное дерево позволяет ответить на запросы «является ли строка t суффиксом s» и «является ли строка t префиксом суффикса, то есть подстрокой s» за время $O(|t|)$. Также оно позволяет получить следующую информацию о строке s и ее подстроках:

– количество различных подстрок строки s. Если спуститься в суффиксном дереве по пути, соответствующему подстроке, мы окажемся либо в вершине, либо посередине ребра (то есть будет пройдена только часть подстроки, соответствующей ребру). Количество различных подстрок s равно количеству различных позиций внутри суффиксного дерева, или сумме длин подстрок, написанных на ребрах, плюс один (положение в корне – пустая подстрока);

– длину наибольшего общего префикса для двух подстрок строки s. Общему префиксу двух строк соответствует общий участок двух путей, идущих от корня;

– лексикографический порядок суффиксов строки s. Обход суффиксного дерева, который в каждой вершине перебирает исходящие ребра в лексикографическом порядке первого символа ребра, перебирает

позиции в дереве в порядке возрастания строк. Отсюда следует, что порядок обхода листьев (позиций, соответствующих суффиксам) есть их лексикографическая сортировка.

Неявное суффиксное дерево строки s – дерево, полученное из суффиксного дерева $s\$$ удалением всех вхождений терминального символа $\$$ из меток ребер дерева, удалением после этого ребер без меток и затем удалением вершин, имеющих меньше двух потомков. Неявное суффиксное дерево префикса $s[1...i]$ строки s получается аналогично из суффиксного дерева для $s[1...i]\$$ удалением символов $\$$, дуг и вершин.

Неявное суффиксное дерево для любой строки s будет иметь меньше листьев, чем суффиксное дерево для строки $s\$$, в том и только том случае, если хотя бы один из суффиксов s является префиксом другого суффикса. Терминальный символ $\$$ добавлен к s как раз во избежание этой ситуации. Если s заканчивается символом, который больше нигде в s не появляется, то неявное суффиксное дерево для s будет иметь лист для каждого суффикса и, следовательно, будет настоящим суффиксным деревом.

Хотя неявное суффиксное дерево может иметь листья не для всех суффиксов, в нем закодированы все суффиксы s – каждый произносится символами какого-либо пути от корня этого неявного суффиксного дерева. Однако если этот путь не кончается листом, то не будет маркера, обозначающего конец пути. Таким образом, неявные суффиксные деревья сами по себе неинформативны.

Обобщенное суффиксное дерево набора строк $s_1 \dots s_n$ – суффиксное дерево, содержащее все суффиксы каждой из n строк. При построении такого дерева каждая строка должна дополняться уникальным символом маркера вне алфавита (или строкой), чтобы гарантировать, что суффикс не является подстрокой другого и представлен уникальным конечным узлом.

Для реализации поставленной задачи выявления вторжений на основе базы сигнатур создается обобщенное суффиксное дерево для сигнатур атак. В данном исследовании за основу взят алгоритм Укконена построения обобщенного суффиксного дерева и обеспечивающий приемлемую сложность [33]. Базовый алгоритм последовательно строит суффиксное дерево для всех префиксов исходного текста $S = s_1 s_2 \dots s_n$. На i-ом шаге неявное суффиксное дерево τ_{i-1} для префикса $s[1...i-1]$ достраивается до τ_i для префикса $s[1...i]$. Для этого

для каждого суффикса подстроки $s[1...i-1]$ выполняют спуск из корня дерева до конца суффикса и дописывают символ s_i. Алгоритм состоит из n этапов, на каждом их которых происходит продление всех суффиксов текущего префикса строки, что требует $O(n^2)$ времени. Общая асимптотика алгоритма – $O(n^3)$.

Пусть $x\alpha$ обозначает произвольную строку, где x – ее первый символ, а α – оставшаяся (возможно пустая) подстрока. Если для внутренней вершины v с путевой меткой $x\alpha$ существует другая вершина $s(v)$ с путевой меткой α, то ссылка из v в $s(v)$ называется суффиксной ссылкой. Для любой внутренней вершины v суффиксного дерева существует суффиксная ссылка, ведущая в некоторую внутреннюю вершину u (пример работы суффиксных ссылок представлен в [34]).

Рассмотрим применение суффиксных ссылок. Пусть только что был продлен суффикс $s[j...i-1]$ до суффикса $s[j...i]$. Теперь с помощью построенных суффиксных ссылок можно найти конец суффикса $s[j+1...i-1]$ в суффиксном дереве, чтобы продлить его до суффикса $s[j+1...i]$. Для этого проходят вверх по дереву до ближайшей внутренней вершины v, в которую ведет путь $s[j...r]$.

У вершины v всегда есть суффиксная ссылка, ведущая к вершине u, которой соответствует путь $s[j+1...r]$. Далее от вершины u следует пройти вниз по дереву к концу суффикса $s[j+1...i-1]$ и продлить его до суффикса $s[j+1...i]$. При этом подстрока $s[j+1...i-1]$ является суффиксом подстроки $s[j...i-1]$. Следовательно, после перехода по суффиксной ссылке в вершину, помеченную путевой меткой $s[j+1...r]$, можно дойти до места, которому соответствует метка $s[r+1...i-1]$, сравнивая не символы на ребрах, а лишь длину ребра по первому символу рассматриваемой части подстроки и длину самой этой подстроки.

В процессе построения суффиксного дерева уже построенные суффиксные ссылки никак не изменяются. Поэтому рассмотрим построение суффиксных ссылок для созданных вершин. Возьмем новую внутреннюю вершину v, которая была создана в результате продления суффикса $s[j...i-1]$. Вместо того, чтобы искать, куда должна указывать

суффиксная ссылка вершины v, проходя путь до корня дерева, продлим следующий суффикс $s[j+1...i-1]$. В этот момент можно проставить суффиксную ссылку для вершины v. Она будет указывать либо на существующую вершину, если следующий суффикс закончился в ней, либо на новую созданную. Таким образом, для вершины v точно найдется на следующем шаге алгоритма внутренняя вершина, в которую должна вести суффиксная ссылка.

Глубиной $d(v)$ вершины v является число ребер на пути от корня до вершины. При переходе по суффиксной ссылке глубина уменьшается не более чем на единицу. Число переходов по ребрам внутри фазы i – $O(i)$. В начале каждой фазы выполняется только один спуск от корня, а затем используются переходы по суффиксным ссылкам. Переходов внутри фазы алгоритма – $O(i)$. Фаза алгоритма состоит из i итераций, и кумулятивно получаем, что на одной итерации будет выполнено $O(1)$ действий. Асимптотика улучшенного алгоритма – $O(n^2)$.

Для дальнейшей оптимизации алгоритма до уровня $O(n)$ предлагается использовать линейное количество памяти. Метку каждого ребра будем сохранять как два числа – позиции ее самого левого и самого правого символов в исходном тексте. Если в какой-то момент работы алгоритма создан лист с меткой i (для суффикса, начинающегося в позиции i строки s), он останется листом во всех последовательных деревьях, созданных алгоритмом. Если правило продления применяется в продолжении суффикса, начинающего в позиции j, оно же и будет применяться во всех дальнейших продолжениях (от $j+1$ до i) до конца фазы алгоритма. Следовательно, в каждой фазе i алгоритм работает с суффиксами из диапазона $j...k, k \leqslant i$ вместо диапазона 1...i.

Алгоритм позволяет обнаруживать с помощью суффиксных деревьев исключительно те последовательности атак, точные копии которых занесены в базу сигнатур. Мутация анализируемой последовательности команд, составляющих атаку, приведут к ошибке второго рода. Данная проблема решена путем разбиения последовательностей на меньшие блоки, что избавляет от влияния мутационных вставок, пропусков и смены порядка команд, так как в случае мутаций позволяет восстанавливать из последовательности одну из наиболее близких сигнатур, находящихся в базе сигнатур.

Обозначим анализируемую последовательность команд $a = C_{a1}C_{a2}C_{a3}\ldots C_{an}$, а известные сигнатуры: $c_1 = C_{c_1,1}C_{c_1,2}C_{c_1,3}\ldots C_{c_1,n_1}$, $c_2 = C_{c_2,1}C_{c_2,2}C_{c_2,3}\ldots C_{c_2,n_2}$, $c_m = C_{c_m,1}C_{c_m,2}C_{c_m,3}\ldots C_{c_m,n_m}$. Каждый вектор атаки из известного множества атак разбивается на пересекающиеся блоки длиной k. На полученных участках строится суффиксное дерево. Для этого очередную сигнатуру $c_i = C_{c_i,1}C_{c_i,2}C_{c_i,3}\ldots C_{c_i,n_i}$ разделим на блоки длины k: $C_{c_i,1}C_{c_i,2}\ldots C_{c_i,k}$; $C_{c_i,2}C_{c_i,3}\ldots C_{c_i,k+1}$; $C_{c_i,n_i-k}C_{c_i,n_i-k+1}\ldots C_{c_i,n_i}$ и добавим полученные подпоследовательности в имеющееся суффиксное дерево.

При сравнении поступившей последовательности a с сигнатурами, по которым уже построено суффиксное дерево, ее аналогично разобьем на блоки длиной k, после чего проведем поиск каждого отдельного блока в суффиксном дереве и вычислим долю совпадений. Если эта доля превышает установленное пороговое значение, то последовательность a отнесем к атаке. Обучение итогового алгоритма сводится к подбору оптимального значения длины k и порога срабатывания.

Рассмотрим пример работы предложенного алгоритма. Пусть в базе сигнатур заданы следующие шаблоны атак:

(1) open, read, open, write, execute, connect, write, write, connect, execute;
(2) execute, execute, connect, open, write, execute, connect, open, read, write.

Построим суффиксное дерево на блоках, выделенных из заданных сигнатур при $k = 3$ (рис. 2, команды обозначены первым символом):

open, read, open; read, open, write; open, write, execute; write, execute, connect; execute, connect, write; connect, write, write; write, write, connect; write, connect, execute.

Разобьем на блоки вторую сигнатуру и добавим их в суффиксное дерево (рис. 3):

execute, execute, connect; execute, connect, open; connect, open, write; open, write, execute; write, execute, connect; execute, connect, open; connect, open, read; open, read, write.

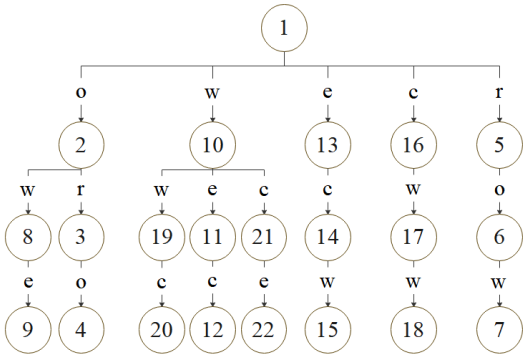


Рис. 2. Суффиксное дерево для первого вектора атаки

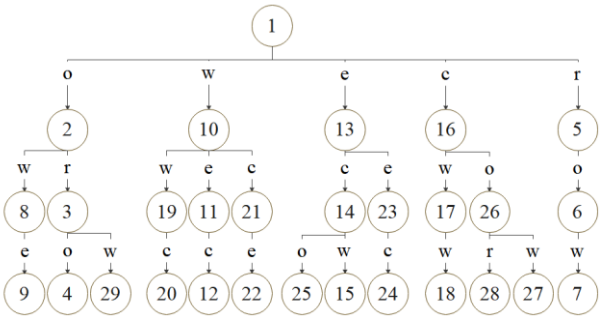


Рис. 3. Суффиксное дерево, дополненное вторым вектором атак

Проанализируем следующую поступившую последовательность на схожесть относительно базы сигнатур:

connect, open, write, execute, connect, write, open, write, execute, connect.

Разобьем данный вектор на блоки той же длины $k = 3$:

connect, open, write; open, write, execute; write, execute, connect; execute, connect, write; connect, write, open; write, open, write; open, write, execute; write, execute, connect.

Поиск в суффиксном дереве показывает, что 6 из 8 блоков (75%) содержатся в дереве. Участки connect, write, open и write, open, write в дереве не содержатся. При пороге 70% поступивший вектор определен как атака.

4. Экспериментальное исследование. Задача экспериментальной оценки созданного метода – оценка уровня ошибок и ресурсных затрат алгоритма при работе на тестовых наборах данных. Для исследований в сетевом режиме использован тестовый набор данных KDD Cup 1999 [35], представляющий собой размеченный датасет последовательностей сетевых пакетов. Использованы наборы для следующих сетевых атак: back dos; multihop r2l; satan probe; buffer_overflow u2r; neptune dos; smurf dos; ftp_write r2l; nmap probe; spy r2l; guess_passwd r2l; perl u2r; teardrop dos; imap r2l; phf r2l; warezclient r2l; ipsweep proe; pod dos; warezmaster r2l; land dos; portsweep probe; loadmodule u2r; rootkit u2r). Испытательный стенд развернут на платформе Intel Core i7, 32 Гб, SSD 1 Тб, операционная система MS Windows 10. Алгоритм выявления вторжений на базе суффиксных деревьев реализован на языке C.

График зависимости уровня ошибок при обнаружении атак от размера базы сигнатур представлен на рисунке 5 (приведен пример для параметров: размер блока $k = 4$, порог срабатывания – 80%.). С ростом размера базы сигнатур незначительно, в пределах десятых долей процента, увеличивается вероятность ошибки первого рода, не превышающая 1%. Вероятность ошибок второго рода при этом сокращается до уровня 0,08.

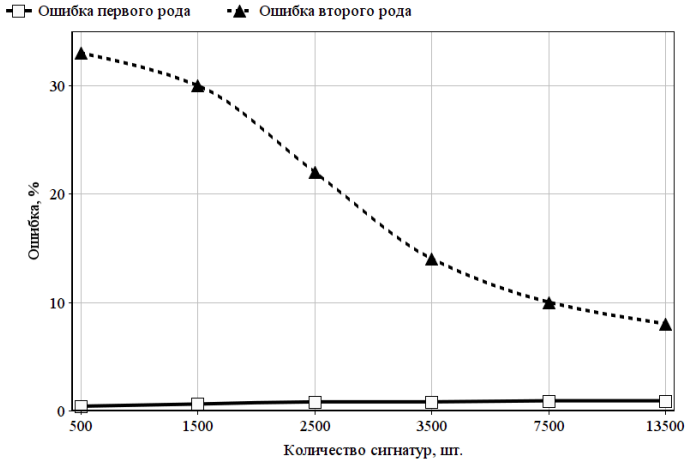


Рис. 5. Оценка уровня ошибок алгоритма ($k = 4$, порог – 80%)

Уровень ресурсозатрат суффиксного алгоритма в зависимости от размера базы сигнатур представлен на рисунках 6 и 7.

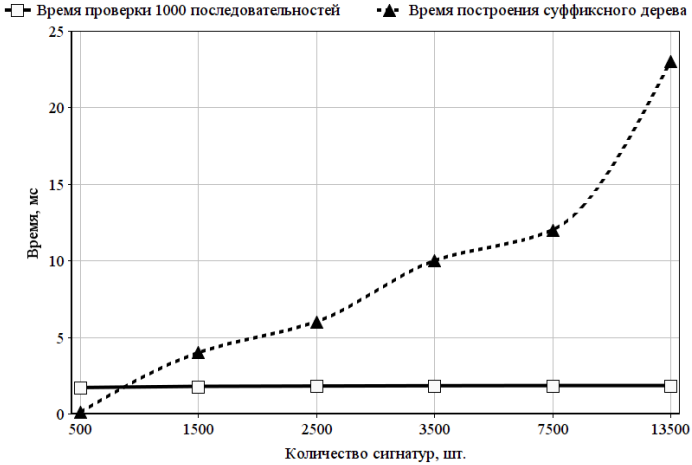


Рис. 6. Оценка временных затрат на проверку
и построение суффиксного дерева

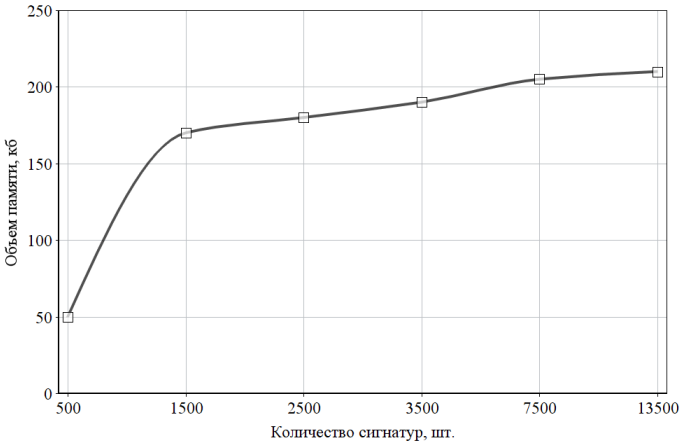


Рис. 7. Оценка объема памяти для хранения суффиксного дерева

Время, затрачиваемое на проверку последовательностей, и объем памяти, занимаемый суффиксным деревом, слабо меняются с наполнением базы сигнатур, что определяется ограниченной глубиной суффиксного дерева, в котором производится поиск, и детерминированностью переходов внутри дерева.

Время построения суффиксного дерева имеет почти линейную зависимость от размеров данных, на которых дерево строится, что подтверждает теоретические оценки алгоритма, приведенные ранее. Повторение одних и тех же элементов векторов и их блоков внутри базы сигнатур позволяет суффиксному дереву компактно хранить данные и эффективно использовать ресурс памяти.

Уровень ошибок второго рода в зависимости от доли мутируемой части анализируемых векторов приведен на рисунке 8 для базы сигнатур размером 13500 записей для тестового набора KDD Cup 1999 с искусственным внесением мутаций (диапазон доли мутаций – 0…30%).

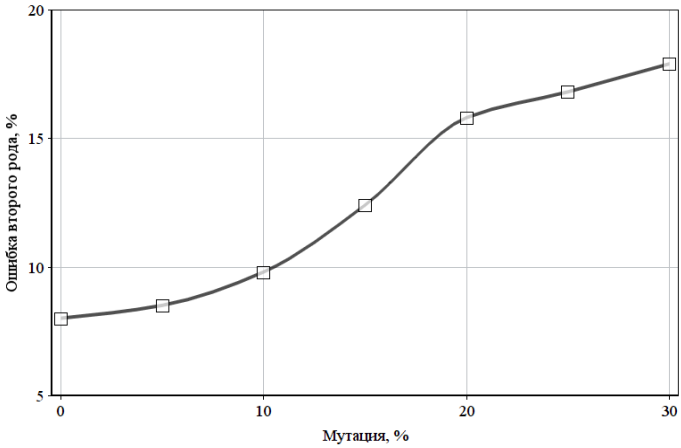


Рис. 8. Влияние мутаций на уровень ошибок алгоритма

Для дальнейшего улучшения алгоритма был проведен ряд модификаций, вследствие которых точность обнаружения атак увеличена до 95% при уровне мутаций 10%. Модифицированный алгоритм отличается по своей эффективности от исходного, комбинируя две системы деревьев, построенные на сигнатурах аномального и нормального поведения. Используется второе суффиксное дерево, построенное на известных шаблонах нормальных (не содержащих атаку) последовательностей. На этапе проверки, даже если проверяемая последовательность имеет высокие показатели совпадения по первому суффиксному дереву, построенному по сигнатурам атак, она не идентифицируется как атака до тех пор, пока коэффициент совпадения со вторым деревом не превысит определенный порог, заданный для второго дерева. Так, например, до модификации при $k = 3$, пороге для первого дерева 65%, пороге для

второго дерева 90% одиночное первое дерево является низкоэффективным, так как ввиду низкого порога система относит подавляющее число векторов к атакам (ошибки первого рода – около 90%), а при добавлении второго дерева ошибка снижается до 2%. При этом тандем суффиксных деревьев замедляет скорость работы алгоритма в N раз, где N – отношение количества векторов нормального поведения к количеству векторов атак в обучающей выборке, но из-за начальной линейной сложности данное замедление можно считать несущественным.

Для сравнения разработанного метода с традиционными СОВ использовались СОВ Suricata и Snort. В условиях отсутствия мутаций в сравниваемых последовательностях СОВ и разработанное решение демонстрируют идентичные показатели точности обнаружения атак. В случае распознавания мутирующих атак разработанное решение сохраняет устойчивость к изменениям в последовательностях. Суффиксный алгоритм позволяет пропускать несовпадающие элементы, и, следовательно, усовершенствовать сигнатурные СОВ, предоставив им новую возможность распознавать атаки, сигнатуры которых явно не содержатся в базе сигнатур. Помимо компактного использования ресурсов памяти для хранения больших баз сигнатур, метод обнаружения вторжений с помощью суффиксных деревьев обладает уникальным свойством – возможностью динамически расширять базу сигнатур во время работы СОВ. Сложность добавления новой ветви суффиксного дерева линейно зависит от длины добавляемого вектора, поэтому такое действие не сказывается на производительности и не требует перезаписи всей базы сигнатур. Это позволяет незамедлительно адаптировать СОВ к новым атакам, добавляемым к базе старых сигнатур.

5. Заключение. В исследовании проанализированы существующие решения в области биоинформатики – алгоритмы сборки и сопоставления геномных последовательностей. Рассмотрена их применимость для решения актуальной задачи идентификации мутирующих атак сигнатурными СОВ и выделен биоинформационный алгоритм обработки последовательностей на базе суффиксных деревьев, который позволяет добиться высокой точности обнаружения вторжений на уровне современных СОВ – более 90%, при этом превосходя их по экономичности использования оперативной памяти, быстродействию и устойчивости к возможным мутациям в векторах атак.

Для улучшения показателей точности проведен ряд модификаций разработанного алгоритма на базе суффиксных деревьев, вследствие которых точность обнаружения атак увеличена до 95% при уровне мутаций в последовательности до 10%.

Дальнейшие исследования направлены на адаптацию разработанного алгоритма для маломощных вычислительных устройств и реализацию для Интернета вещей сигнатурной СОВ на базе предложенного решения.

Разработанная технология обнаружения мутирующих атак с помощью суффиксных деревьев и СОВ, использующие его, могут применяться для обнаружения вторжений как в классических компьютерных сетях, так и в современных реконфигурируемых сетевых инфраструктурах с ограниченными ресурсами (Интернет вещей, сети киберфизических объектов, сенсорные сети).

Литература

1. *Khraisat A., Gondal I., Vamplew P., Kamruzzaman J.* Survey of intrusion detection systems: techniques, datasets and challenges // Cybersecurity. 2019. vol. 2. no. 1.
2. *Jatti S.A.V., Kishor Sontif V.J.K.* Intrusion detection systems // International Journal of Recent Technology and Engineering. 2019. vol. 8. no. 2. special issue 11. pp. 3976–3983.
3. *Branitskiy A.A., Kotenko I.V.* Analysis and classification of methods for network attack detection // SPIIRAS Proceedings. 2016. vol. 2. no. 45. pp. 207–244.
4. *Lakshminarayana D.H., Philips J., Tabrizi N.* A survey of intrusion detection techniques // In Proceedings - 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019. 2019. pp. 1122–1129.
5. *Platonov V.V., Semenov P.O.* An adaptive model of a distributed intrusion detection system // Automatic Control and Computer Sciences. 2017. vol. 51. no. 8. pp. 894–898.
6. *Platonov V.V., Semenov P.O.* Detection of Abnormal Traffic in Dynamic Computer Networks with Mobile Consumer Devices // Automatic Control and Computer Sciences, 2018. vol. 52. no. 8. pp. 959–964.
7. *Aljawarneh S.A., Moftah R.A., Maatuk A.M.* Investigations of automatic methods for detecting the polymorphic worms signatures // Future Generation Computer Systems. 2016. vol. 60. pp. 67–77.
8. *Khonde S.R., Venugopal U.* Hybrid architecture for distributed intrusion detection system // Ingenierie des Systemes d'Information. 2019. vol. 24. no. 1. pp. 19–28.
9. *Zhang W.A., Hong Z., Zhu J.W., Chen B.* A survey of network intrusion detection methods for industrial control systems // Kongzhi yu Juece/Control and Decision. 2019. vol. 34. no. 11. pp. 2277–2288.
10. *Seoane Fernández J.A., Miguélez Rico M.* Bio-Inspired Algorithms in Bioinformatics I // Encyclopedia of Artificial Intelligence. 2011.
11. *Levshun D, Gaifulina D., Chechulin A., Kotenko I.* Problematic issues of information security of cyber-physical systems // SPIIRAS Proceedings. 2020. vol. 19. no. 5. pp. 1050–1088.
12. *Coull S., Branch J., Szymanski B., Breimer E.* Intrusion detection: A bioinformatics approach // In Proceedings Annual Computer Security Applications Conference, ACSAC. 2003. vol. 2003-January. pp. 24–33.
13. *Lavrova D., Zaitceva E., Zegzhda P.* Bio-inspired approach to self-regulation for industrial dynamic network infrastructure // CEUR Workshop Proceedings. 2019. vol. 2603. pp. 34–39.
14. *Miller W.* An Introduction to Bioinformatics Algorithms // Journal of the American Statistical Association. 2006. vol. 101. no. 474. pp. 855–855.

15. *Sohn J. Il, Nam J.W.* The present and future of de novo whole-genome assembly // Briefings in Bioinformatics. 2018. vol. 19, no. 1, pp. 23–40.
16. *Recanati A., Brüls T., D'Aspremont A.* A spectral algorithm for fast de novo layout of uncorrected long nanopore reads // Bioinformatics. 2017. vol. 33, no. 20. pp. 3188–3194.
17. *Rizzi R., et al.* Overlap graphs and de Bruijn graphs: data structures for de novo genome assembly in the big data era // Quantitative Biology. 2019. vol. 7, no. 4. pp. 278–292.
18. *Wittler R.* Alignment- And reference-free phylogenomics with colored de Bruijn graphs // Algorithms for Molecular Biology. 2020. vol. 15. no. 1.
19. *Tan T.W., Lee E.* Sequence Alignment // Beginners Guide to Bioinformatics for High Throughput Sequencing. 2018. pp. 81–115.
20. *Muhamad F.N., Ahmad R.B., Asi S.M., Murad M.N.* Performance Analysis of Needleman-Wunsch Algorithm (Global) and Smith-Waterman Algorithm (Local) in Reducing Search Space and Time for DNA Sequence Alignment // Journal of Physics: Conference Series. 2018. vol. 1019. no. 1.
21. *Lee Y.S., Kim Y.S., Uy R.L.* Serial and parallel implementation of Needleman-Wunsch algorithm // International Journal of Advances in Intelligent Informatics. 2020. vol. 6. no. 1. pp. 97–108.
22. *Čavojský M., Drozda M., Balogh Z.* Analysis and experimental evaluation of the Needleman-Wunsch algorithm for trajectory comparison // Expert Systems with Applications. 2021. vol. 165.
23. *Sun J., Chen K., Hao Z.* Pairwise alignment for very long nucleic acid sequences // Biochemical and Biophysical Research Communications. 2018. vol. 502. no. 3. pp. 313–317.
24. *Zou H., Tang S., Yu C., Fu H., Li Y., Tang W.* ASW: Accelerating Smith–Waterman Algorithm on Coupled CPU-GPU Architecture // International Journal of Parallel Programming. 2019. vol. 47. no. 3. pp. 388–402.
25. *Chowdhury B., Garai G.* A review on multiple sequence alignment from the perspective of genetic algorithm // Genomics. 2017. vol. 109. no. 5–6. pp. 419–431.
26. *Dijkstra M.J.J., Van Der Ploeg A.J., Feenstra K. A., Fokkink W.J., Abeln S., Heringa J.* Tailor-made multiple sequence alignments using the PRALINE 2 alignment toolkit // Bioinformatics. 2019. vol. 35. no. 24. pp. 5315–5317.
27. *Chen S., Yang S., Zhou M., Burd R., Marsic I.* Process-Oriented Iterative Multiple Alignment for Medical Process Mining // In IEEE International Conference on Data Mining Workshops, ICDMW. 2017. vol. 2017-November. pp. 438–445.
28. *Ye N.* Markov Chain Models and Hidden Markov Models // Data Mining. 2021. pp. 287–305.
29. *Behera N., Jeevitesh M.S., Jose J., Kant K., Dey A., Mazher J.* Higher accuracy protein multiple sequence alignments by genetic algorithm // Procedia Computer Science. 2017. vol. 108. pp. 1135–1144.
30. *Cui X., Shi H., Zhao J., Ge Y., Yin Y., Zhao K.* High Accuracy Short Reads Alignment Using Multiple Hash Index Tables on FPGA Platform // In Proceedings of 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference, ITOEC. 2020. pp. 567–573.
31. *Marçais G., Delcher A.L., Phillippy A.M., Coston R., Salzberg S.L., Zimin A.* MUMmer4: A fast and versatile genome alignment system // PLoS Computational Biology. 2018. vol. 14. no. 1. 2018.
32. *Kay M.* Substring alignment using suffix trees // Lecture Notes in Computer Science. 2004. vol. 2945. pp. 275–282.
33. *Ukkonen E.* On-line construction of suffix trees // Algorithmica. 1995. vol. 14. no. 3. pp. 249–260.

34. *Breslauer D., Italiano G.F.* On suffix extensions in suffix trees // Theoretical Computer Science. 2012. vol. 457. pp. 27–34.
35. KDD Cup 1999 Data: URL: kdd.ics.uci.edu/databases/kddcup99/kddcup99.html (дата доступа: 10.04.2021).

Зегжда Дмитрий Петрович — д-р техн. наук, профессор РАН, директор института, Институт кибербезопасности и защиты информации, Санкт-Петербургский политехнический университет Петра Великого. Область научных интересов: моделирование безопасности, методы киберустойчивости, технология безопасности систем больших данных. Число научных публикаций — 340. dmitry@ibks.spbstu.ru; Политехническая улица, д. 29, г. Санкт-Петербург, 192251, РФ; р.т.: +7(812)5527632, факс: +7(812)5527632.

Калинин Максим Олегович — д-р техн. наук, профессор, профессор, Институт кибербезопасности и защиты информации, Санкт-Петербургский политехнический университет Петра Великого. Область научных интересов: анализ киберрисков, методы машинного обучения, методы оценки безопасности реконфигурируемых систем на моделях. Число научных публикаций — 320. max@ibks.spbstu.ru; Политехническая улица, д. 29, г. Санкт-Петербург, 192251, РФ; р.т.: +7(812)5527632, факс: +7(812)5527632.

Крундышев Василий Михайлович — аспирант, ассистент, Институт кибербезопасности и защиты информации, Санкт-Петербургский политехнический университет Петра Великого. Область научных интересов: методы искусственного интеллекта в обработке данных о защищенности сложных систем, технологии виртуализации, методы моделирования средств защиты. Число научных публикаций — 60. vmk@ibks.spbstu.ru; Политехническая улица, д. 29, г. Санкт-Петербург, 192251, РФ; р.т.: +7(812)5527632, факс: +7(812)5527632.

Лаврова Дарья Сергеевна — д-р техн. наук, профессор, Институт кибербезопасности и защиты информации, Санкт-Петербургский политехнический университет Петра Великого. Область научных интересов: биоподобные методы кибербезопасности, методы высокопроизводительного анализа больших данных, технологии обработки событий безопасности. Число научных публикаций — 270. lavrova@ibks.spbstu.ru; Политехническая улица, д. 29, г. Санкт-Петербург, 192251, РФ; р.т.: +7(812)5527632, факс: +7(812)5527632.

Москвин Дмитрий Андреевич — к-т техн. наук, доцент, Институт кибербезопасности и защиты информации, Санкт-Петербургский политехнический университет Петра Великого. Область научных интересов: верификация безопасности, методы model checking, эвристические и поведенческие модели безопасности. Число научных публикаций — 120. moskvin@ibks.spbstu.ru; Политехническая улица, д. 29, г. Санкт-Петербург, 192251, РФ; р.т.: +7(812)5527632, факс: +7(812)5527632.

Павленко Евгений Юрьевич — к-т техн. наук, доцент, Институт кибербезопасности и защиты информации, Санкт-Петербургский политехнический университет Петра Великого. Область научных интересов: адаптивные методы управления безопасностью, модели управления и принятия решений, графовые алгоритмы моделирования. Число научных публикаций — 190. pavlenko@ibks.spbstu.ru; Политехническая улица, д. 29, г. Санкт-Петербург, 192251, РФ; р.т.: +7(812)5527632, факс: +7(812)5527632.

Поддержка исследований. Работа выполнена в рамках Государственного задания на проведение фундаментальных исследований (код темы 0784-2020-0026).

DOI 10.15622/ia.20.4.3

D. ZEGZHDA, M. KALININ, V. KRUNDYSHEV, D. LAVROVA,
D. MOSKVIN, E. PAVLENKO

APPLICATION OF BIOINFORMATICS ALGORITHMS FOR POLYMORPHIC CYBERATTACKS DETECTION

Zegzhda D., Kalinin M., Kundyshev V., Lavrova D., Moskvin D., Pavlenko E. **Application of Bioinformatics Algorithms for Polymorphic Cyberattacks Detection.**

Abstract. The functionality of any system can be represented as a set of commands that lead to a change in the state of the system. The intrusion detection problem for signature-based intrusion detection systems is equivalent to matching the sequences of operational commands executed by the protected system to known attack signatures. Various mutations in attack vectors (including replacing commands with equivalent ones, rearranging the commands and their blocks, adding garbage and empty commands into the sequence) reduce the effectiveness and accuracy of the intrusion detection. The article analyzes the existing solutions in the field of bioinformatics and considers their applicability for solving the problem of identifying polymorphic attacks by signature-based intrusion detection systems. A new approach to the detection of polymorphic attacks based on the suffix tree technology applied in the assembly and verification of the similarity of genomic sequences is discussed. The use of bioinformatics technology allows us to achieve high accuracy of intrusion detection at the level of modern intrusion detection systems (more than 0.90), while surpassing them in terms of cost-effectiveness of storage resources, speed and readiness to changes in attack vectors. To improve the accuracy indicators, a number of modifications of the developed algorithm have been carried out, as a result of which the accuracy of detecting attacks increased by up to 0.95 with the level of mutations in the sequence up to 10%. The developed approach can be used for intrusion detection both in conventional computer networks and in modern reconfigurable network infrastructures with limited resources (Internet of Things, networks of cyber-physical objects, wireless sensor networks).

Keywords: Ukkonen Algorithm, Security, Bioinformatics, Alignment, Mutation, Intrusion Detection, Polymorphism, Signature, Suffix Tree.

Zegzhda Dmitry — Dr.Sc., Professor of the Russian Academy of Sciences, Director of Institute, Institute for Cybersecurity and Information Protection, Peter the Great St. Petersburg Polytechnic University. Research interests: security modeling, cyber resilience methods, big data system security technology. The number of publications – 340. dmitry@ibks.spbstu.ru; 29, Politekhnicheskaya ul., St. Petersburg, 192251, Russian Federation; office phone: +7(812)5527632, fax: +7(812)5527632.

Kalinin Maxim — Dr.Sc., Professor, Professor, Institute for Cybersecurity and Information Protection, Peter the Great St. Petersburg Polytechnic University. Research interests: cyber risk analysis, machine learning methods, methods for evaluating the security of reconfigurable systems on models. The number of scientific publications – 320. max@ibks.spbstu.ru; 29, Politekhnicheskaya ul., St. Petersburg, 192251, Russian Federation; office phone: +7(812)5527632, fax: +7(812)5527632.

Krundyshev Vasiliy — Ph.D student, Junior Researcher, Assistant, Institute for Cybersecurity and Information Protection, Peter the Great St. Petersburg Polytechnic University. Research interests: methods of artificial intelligence in the processing of data on the security of complex

systems, virtualization technologies, methods of modeling security tools. The number of scientific publications – 60. vmk@ibks.spbstu.ru; 29, Politekhnicheskaya ul., St. Petersburg, 192251, Russian Federation; office phone: +7(812)5527632, fax: +7(812)5527632.

Lavrova Daria — Dr.Sc., Professor, Institute for Cybersecurity and Information Protection, Peter the Great St. Petersburg Polytechnic University. Research interests: bioinspired methods for cybersecurity, methods of high-performance analysis of big data, security event processing technologies. The number of scientific publications – 270. lavrova@ibks.spbstu.ru; 29, Politekhnicheskaya ul., St. Petersburg, 192251, Russian Federation; office phone: +7(812)5527632, fax: +7(812)5527632.

Moskvin Dmitry — PhD, Associate Professor, Institute for Cybersecurity and Information Protection, Peter the Great St. Petersburg Polytechnic University. Research interests: security verification, model checking methods, heuristic and behavioral security models. The number of scientific publications – 120. moskvin@ibks.spbstu.ru; 29, Politekhnicheskaya ul., St. Petersburg, 192251, Russian Federation; office phone: +7(812)5527632, fax: +7(812)5527632.

Pavlenko Evgeny — PhD, Associate Professor, Institute for Cybersecurity and Information Protection, Peter the Great St. Petersburg Polytechnic University. Research interests: adaptive methods of security management, models of management and decision making, graph modeling algorithms. The number of scientific publications – 190. pavlenko@ibks.spbstu.ru; 29, Politekhnicheskaya ul., St. Petersburg, 192251, Russian Federation; office phone: +7(812)5527632, fax: +7(812)5527632.

Acknowledgements. The work was performed as part of the State assignment for basic research (topic code 0784-2020-0026).

References

1. Khraisat A., Gondal I., Vamplew P., Kamruzzaman J. Survey of intrusion detection systems: techniques, datasets and challenges. *Cybersecurity*. 2019. vol. 2. no. 1.
2. Jatti S.A.V., Kishor Sontif V.J.K. Intrusion detection systems. *International Journal of Recent Technology and Engineering*. 2019. vol. 8. no. 2. special issue 11. pp. 3976–3983.
3. Branitskiy A.A., Kotenko I.V. Analysis and classification of methods for network attack detection. *SPIIRAS Proceedings*. 2016. vol. 2. no. 45. pp. 207–244.
4. Lakshminarayana D.H., Philips J., Tabrizi N. A survey of intrusion detection techniques. *In Proceedings - 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019*. 2019. pp. 1122–1129.
5. Platonov V.V., Semenov P.O. An adaptive model of a distributed intrusion detection system. *Automatic Control and Computer Sciences*. 2017. vol. 51. no. 8. pp. 894–898.
6. Platonov V.V., Semenov P.O. Detection of Abnormal Traffic in Dynamic Computer Networks with Mobile Consumer Devices. *Automatic Control and Computer Sciences*, 2018. vol. 52. no. 8. pp. 959–964.
7. Aljawarneh S.A., Moftah R.A., Maatuk A.M. Investigations of automatic methods for detecting the polymorphic worms signatures. *Future Generation Computer Systems*. 2016. vol. 60. pp. 67–77.
8. Khonde S.R., Venugopal U. Hybrid architecture for distributed intrusion detection system. *Ingenierie des Systemes d'Information*. 2019. vol. 24. no. 1. pp. 19–28.
9. Zhang W.A., Hong Z., Zhu J.W., Chen B. A survey of network intrusion detection methods for industrial control systems. *Kongzhi yu Juece/Control and Decision*. 2019. vol. 34. no. 11. pp. 2277–2288.

10. Levshun D, Gaifulina D., Chechulin A., Kotenko I. Problematic issues of information security of cyber-physical systems. *SPIIRAS Proceedings*. 2020. vol. 19. no. 5. pp. 1050–1088.
11. Seoane Fernández J.A., Miguélez Rico M. Bio-Inspired Algorithms in Bioinformatics I. *Encyclopedia of Artificial Intelligence*, 2011.
12. Coull S., Branch J., Szymanski B., Breimer E. Intrusion detection: A bioinformatics approach. In Proceedings Annual Computer Security Applications Conference, ACSAC. 2003. vol. 2003-January. pp. 24–33.
13. Lavrova D., Zaitceva E., Zegzhda P. Bio-inspired approach to self-regulation for industrial dynamic network infrastructure. CEUR Workshop Proceedings. 2019. vol. 2603. pp. 34–39.
14. Miller W. An Introduction to Bioinformatics Algorithms. *Journal of the American Statistical Association*. 2006. vol. 101. no. 474. pp. 855–855.
15. Sohn J. Il, Nam J.W. The present and future of de novo whole-genome assembly. *Briefings in Bioinformatics*. 2018. vol. 19. no. 1. pp. 23–40.
16. Recanati A., Brüls T., D'Aspremont A. A spectral algorithm for fast de novo layout of uncorrected long nanopore reads. *Bioinformatics.* 2017. vol. 33, no. 20. pp. 3188–3194.
17. Rizzi R., et al. Overlap graphs and de Bruijn graphs: data structures for de novo genome assembly in the big data era. *Quantitative Biology*. 2019. vol. 7. no. 4. pp. 278–292.
18. Wittler R. Alignment- And reference-free phylogenomics with colored de Bruijn graphs. *Algorithms for Molecular Biology*. 2020. vol. 15. no. 1.
19. Tan T.W., Lee E. Sequence Alignment. *In Beginners Guide to Bioinformatics for High Throughput Sequencing*. 2018. pp. 81–115.
20. Muhamad F.N., Ahmad R.B., Asi S.M., Murad M.N. Performance Analysis of Needleman-Wunsch Algorithm (Global) and Smith-Waterman Algorithm (Local) in Reducing Search Space and Time for DNA Sequence Alignment. *Journal of Physics: Conference Series.* 2018. vol. 1019. no. 1.
21. Lee Y.S., Kim Y.S., Uy R.L. Serial and parallel implementation of Needleman-Wunsch algorithm. *International Journal of Advances in Intelligent Informatics*. 2020. vol. 6. no. 1. pp. 97–108.
22. Čavojský M., Drozda M., Balogh Z. Analysis and experimental evaluation of the Needleman-Wunsch algorithm for trajectory comparison. *Expert Systems with Applications*. 2021. vol. 165.
23. Sun J., Chen K., Hao Z. Pairwise alignment for very long nucleic acid sequences. *Biochemical and Biophysical Research Communications*. 2018. vol. 502. no. 3. pp. 313–317.
24. Zou H., Tang S., Yu C., Fu H., Li Y., Tang W. ASW: Accelerating Smith–Waterman Algorithm on Coupled CPU-GPU Architecture. *International Journal of Parallel Programming*. 2019. vol. 47. no. 3. pp. 388–402.
25. Chowdhury B., Garai G. A review on multiple sequence alignment from the perspective of genetic algorithm. *Genomics.* 2017. vol. 109. no. 5–6. pp. 419–431.
26. Dijkstra M.J.J., Van Der Ploeg A.J., Feenstra K. A., Fokkink W.J., Abeln S., Heringa J. Tailor-made multiple sequence alignments using the PRALINE 2 alignment toolkit. *Bioinformatics.* 2019. vol. 35. no. 24. pp. 5315–5317.
27. Chen S., Yang S., Zhou M., Burd R., Marsic I. Process-Oriented Iterative Multiple Alignment for Medical Process Mining. In IEEE International Conference on Data Mining Workshops, ICDMW. 2017. vol. 2017-November. pp. 438–445.
28. Ye N. Markov Chain Models and Hidden Markov Models. *Data Mining.* 2021. pp. 287–305.
29. Behera N., Jeevitesh M.S., Jose J., Kant K., Dey A., Mazher J. Higher accuracy protein multiple sequence alignments by genetic algorithm. *Procedia Computer Science.* 2017. vol. 108. pp. 1135–1144.

30. Cui X., Shi H., Zhao J., Ge Y., Yin Y., Zhao K. High Accuracy Short Reads Alignment Using Multiple Hash Index Tables on FPGA Platform. In Proceedings of IEEE 5th Information Technology and Mechatronics Engineering Conference, ITOEC. 2020. pp. 567–573.
31. Marçais G., Delcher A.L., Phillippy A.M., Coston R., Salzberg S.L., Zimin A. MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology*. 2018. vol. 14. no. 1.
32. Kay M. Substring alignment using suffix trees. *Lecture Notes in Computer Science*. 2004. vol. 2945. pp. 275–282.
33. Ukkonen E. On-line construction of suffix trees. *Algorithmica*. 1995. vol. 14. no. 3. pp. 249–260.
34. Breslauer D., Italiano G.F. On suffix extensions in suffix trees. Theoretical Computer Science. 2012. vol. 457, pp. 27–34.
35. KDD Cup 1999 Data. Available at: kdd.ics.uci.edu/databases/kddcup99/kddcup99.html (accessed: 10.04.2021).

DOI 10.15622/ia.20.4.4

Y. CHEVALIER, F. FENZL, M. KOLOMEETS, R. RIEKE, A. CHECHULIN, C. KRAUSS

CYBERATTACK DETECTION IN VEHICLES USING CHARACTERISTIC FUNCTIONS, ARTIFICIAL NEURAL NETWORKS AND VISUAL ANALYSIS

Chevalier Y., Fenzl F., Kolomeets M., Rieke R., Chechulin A., Krauss C. **Cyberattack detection in vehicles using characteristic functions, artificial neural networks and visual analysis.**

Abstract. The connectivity of autonomous vehicles induces new attack surfaces and thus the demand for sophisticated cybersecurity management. Thus, it is important to ensure that in-vehicle network monitoring includes the ability to accurately detect intrusive behavior and analyze cyberattacks from vehicle data and vehicle logs in a privacy-friendly manner. For this purpose, we describe and evaluate a method that utilizes characteristic functions and compare it with an approach based on artificial neural networks. Visual analysis of the respective event streams complements the evaluation. Although the characteristic functions method is an order of magnitude faster, the accuracy of the results obtained is at least comparable to those obtained with the artificial neural network. Thus, this method is an interesting option for implementation in in-vehicle embedded systems. An important aspect for the usage of the analysis methods within a cybersecurity framework is the explainability of the detection results.

Keywords: controller area network security, intrusion detection, anomaly detection, machine learning, automotive security, security monitoring.

1. Introduction. Information technology (IT) security and data protection are essential for the Internet of Vehicles [1]. Due to a strong connectivity of vehicles and the dependency on external information sources and services, the attack surface increases for intelligent autonomous vehicles. This is exacerbated by the increasing complexity of modern vehicles with more than 100 electronic control units (ECUs) and more than 100 million lines of code. Thus, vulnerabilities in software are highly likely that could be exploited by a potential attacker. However, it is imperative that an attacker cannot influence safety-critical systems. But it has already been demonstrated in [2] how an attacker can remotely take over ECUs to influence steering and braking. It is therefore very important to improve the security of in-vehicle networks, and as long as there are no effective means to prevent certain attacks, methods should be in place to automatically detect them and respond accordingly. The United Nations Economic Commission for Europe (UNECE) has issued Regulation No. 155, "Cybersecurity and Cybersecurity Management System- [3], which makes cybersecurity mandatory for the approval of new vehicle types. One important requirement is that vehicle manufacturers must implement mechanisms to detect cyberattacks in a privacy-friendly manner. This includes measures to detect denial-of-service attacks, such as when the

Controller Area Network (CAN) bus is flooded or ECUs are crashed by a high message load, and subsequent recovery. Furthermore measures for the detection of malicious messages need to be implemented. In principle, the detection of anomalies in network traffic within a vehicle caused by attackers could be done remotely by sending all internal traffic to a Security Operation Center (SOC). However, this would be problematic from a privacy perspective, it would be inefficient, it would incur high costs, and it might not meet real-time response requirements.

This paper is based on our own preliminary work presented in [4]. We propose a new method for in-vehicle anomaly detection that satisfies the following four requirements: (1) the recognition accuracy should be equivalent to or better than existing IDS systems, (2) the method should be lightweight and resource efficient so that it can be executed on typical ECUs, (3) no hardware changes should be necessary and no additional third-party software libraries should be required (as they may not be available for specific ECUs), and (4) the anomaly detection results should be explainable in order to make informed decisions about countermeasures.

To meet these requirements, we propose a logic analysis method that we compare to an artificial neural network-based method that could likely be used in embedded systems in vehicles. We aim for better accuracy, faster and more resource-efficient message characterization, portability to embedded systems without dependencies on libraries such as Tensorflow, and rule-based reasoning so that message evaluation results related to anomalies can be traced back to the responsible rules. We evaluate the proposed method on data sets of the CAN bus, which is the standard solution for communication between ECUs in vehicles.

The remainder of this paper is organized as follows: Section 2 gives an overview on the background and related work. Section 3 introduces data sets from two different vehicles that have been used to evaluate the proposed method. Section 4 presents the principles of the characteristic functions method while Section 5 describes its implementation and the results of various detection setups. Section 6 describes some results from tests with neural networks in order to provide a benchmark for our work. Finally, Section 7 concludes this paper.

2. Background and Related Work. The security of a system can be improved by reducing its attack surface. In [8, 9], for example, possible break-in points are listed together with suggestions for countermeasures such as cryptography, detection of anomalies and ensuring software integrity by separating critical systems. However, most of the intrusion prevention measures currently under discussion require hardware changes, which is inconsistent

with backward compatibility. Therefore, researchers and industry experts have suggested that in the CAN context intrusion detection should be used in addition to the established security mechanisms [10, 11]. CAN intrusion detection methods can be divided into four categories: ECU imitation detection, specification violation detection, message insertion detection, and sequence context anomaly detection. The work to recognize ECU imitations like [12, 13] uses in most cases some kind of physical fingerprint through voltage or time analysis with specific hardware. This work tries to alleviate the general problems of missing authenticity measures in the CAN bus design and thus complements the work presented in this paper. Specification violation detection requires the normal behavior specification to be available, and therefore the benefit of not generating warnings based on false positives. Specification-based intrusion detection methods can use specific checks, e.g. for formality, protocol and data area [14], a certain frequency sensor [15], a number of network-based detection sensors [16] or specifications of the state machines [17]. Message insertion detection can be based on various technologies, such as the analysis of time intervals of messages [18] or long short-term memory (LSTM) [19]. The methods for detecting sequence context anomalies include process mining [20], hidden Markov models [21, 22], a one class Support Vector Machine (OCSVM) [23], artificial neural networks [24] and detection of anomalous patterns in a transition matrix [25]. In most cases, the authors of the above papers have described experiments with a specific method. However, since the authors use different data sets for their experiments, the results of their work cannot be directly compared. Comparisons of various machine learning (ML) algorithms are included in [6, 26, 27]. OCSVM, Self Organizing Maps, and LSTM are used in [26] while LSTM, Gated Recurrent Units (GRU) and Markov models are used in [27]. OCSVM, SVM, sequential neural networks and LSTM are used in [6]. A detailed overview on intrusion detection systems for in-vehicle networks can be found in [28].

The method of characteristic functions used here has already proven in [4] to be much faster and more resource-efficient than artificial neural network methods. With respect to our previous work presented in [4] which was using training sets from [5] and [6], we now use improved state-of-art log files from [7] which have been designed to include sophisticated attack types which do not disrupt normal timing, and thus would not be detected with a frequency-based intrusion detection system (IDS).

3. In-Vehicle Attacks and Data Sets. To evaluate our work, we used the Oak Ridge National Laboratory's (ORNL) Road data set, which was originally presented in [7]. The data set is presented in form of a set of text

files from which 4 meaningful fields can be extracted, that are then mapped in the structure that is presented in Table 1:

1. *Time*. Capture time – helps to identify the time sequence of messages.
2. *ID*. CAN ID, where the lowest ID has a higher channel priority.
3. *Payload*. $p_1, \ldots, p_8$ – 8 bytes with data.
4. *Type*. Attack label where -1 is attack and 1 is the legal message.

The original data set does not contain ground truth flags for each message, but metadata files that describe the attack in regards to timing, content and target ID. For our method we reformatted the messages and marked each valid message in the log with a `1` and each malicious intrusion message with a `-1`. In the exemplary excerpt of a data log in Table 1 you can see that the first occurrence of a message with arbitration ID `208` is a valid message sent by the responsible ECU, whereas the second occurrence is a malicious message introduced by an intruder.

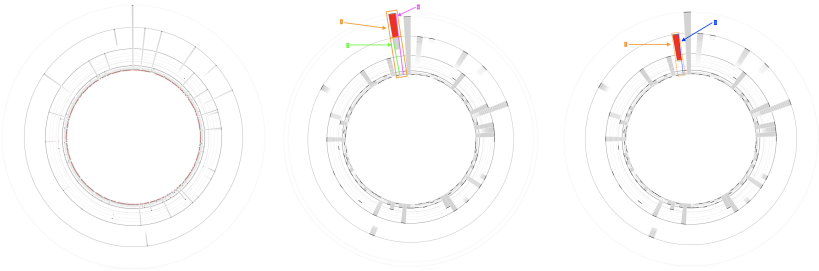
Table 1. Exemplary messages from the ORNL Road data set intrusion scenario `max_speedometer_attack_1`

i	Time	ID	p^1	p^2	p^3	p^4	p^5	p^6	p^7	p^8	Type
1	42.0256	1533	189	221	253	128	126	255	237	218	1
2	42.0271	208	10	115	4	100	136	5	110	0	1
3	42.0282	51	0	6	128	0	12	66	183	208	1
4	42.0282	263	0	0	0	0	0	0	0	0	1
5	42.0282	4095	0	0	0	0	0	0	0	0	1
6	42.0282	14	32	82	150	2	8	9	118	148	1
7	42.0292	208	10	115	4	100	136	255	110	0	-1
8	42.0292	293	144	0	65	31	64	255	163	96	1
9	42.0292	186	6	152	5	4	16	0	2	100	1

Attacks that are presented in the ORNL data set can be divided into 3 categories:

1. *Fuzzing attack* – an attacker injects messages with maximum payloads for many random CAN IDs.
2. *Message injection target attack* – an attacker inject message with a specific CAN ID immediately after the legal message appeared. Thus injected messages are superimposed on legal messages.
3. *Message injection target attack with masquerade* – this attack is similar to the previous one, but legitimate messages were removed. So injected messages replace legal ones.

For a better understanding of what traffic looks like with injected and legal messages, we present these 3 attacks using radial bar chart visualization that was originally presented in [29]. Examples of visualization of these 3 attack types are presented in Figure 1, where malicious messages are marked with red bar color and red bubble. In this visualization, we present attack



(a) Fuzzing attack (b) Message injection attack (c) Masquerade attack

Fig. 1. Different attack scenarios in ORNL Road data set, where red bars and bubbles indicate injected messages and grey bars indicate legal messages

influence by radial time intervals, where each CAN ID is represented as a bar whose height equals the number of messages. Bars consist of arcs that represent payload — the more messages with the same payload the higher is the arc. So solid (or almost solid) arcs depict messages with the same payload, while thin or even transparent (their thickness is less than a pixel) bars depict messages with big payload variety.

Fuzzing attack is the simplest for detection using visual analytics. Usually Fuzzy attack is characterized by many CAN IDs with almost empty bars (see red bubbles in Fig.1a).

For message injection attack there is a pattern in the radial chart – CAN ID with injection have 2 types of message frequency distribution. The first distribution that is without injected messages consists of thin arcs with various payloads. The distribution of injected messages is a solid bar that indicates a lack of variability of payload. We can see how injected messages are superimposed on legal ones by the next patterns depicted in Figure 1b:

1. Sharp difference between frequency (orange indicator #1) – the first part of the bar is very frequent (legal messages) and the second one is not (injected messages).

2. A part of such bar (legal messages – green indicator #2) has almost the same number of messages that other bars.

3. Whole bar (legal messages plus injected – purple indicator #3) does not have the same number of messages that other bars.

For masquerade attack the injection is not clearly detectable by visualization (see Fig.1c). The bars with injected messages look pretty normal except (orange indicator #1) sharp difference between frequency where legal messages have various payload and injected do not. But the (blue indicator #2)

height of the bar looks normal, as the attacker replaced the legal messages with malicious ones.

The ORNL data set also contains "Accelerator Attacks"data sets. This type of attack exploits a vulnerability that puts the ECUs into a compromised state. Therefore, there are no injected messages, so we do not analyze "Accelerator Attacks"data sets in this paper.

4. Principles of the Characteristic Functions Method. Before proceeding to the presentation of characteristic functions, we start by formalising a generic notion of intrusion detection in Section 4.1, and prove that this setting is not usable for exhaustive search in practice. We then present in Section 4.2 criteria that we have considered, and the characteristic functions in Section 4.3

4.1. Formal setting. To formalize this notion, we need to introduce a few notations. We let $\mathcal{P} = [0, 255]^8$ be the set of CAN messages payloads, $\mathcal{I} \subseteq [0, 4095]$ be the set of CAN messages ID, and $\mathcal{B} = \{\bot, \top\}$ denote the respective false and true values. A *log* of length n is a finite sequence $(e_i)_{1 \leqslant i \leqslant n}$ of elements in $\mathcal{I} \times \mathcal{P}$. Let $\mathcal{L}$ be the set of logs. Given a log $L \in \mathcal{L}$ and $\iota \in \mathcal{I}$, we let $\pi_\iota(L)$ be the subsequence of L of elements whose ID is ι. Given a log $L = (e_i)_{1 \leqslant i \leqslant n}$ of length n and $1 \leqslant k \leqslant n$, we denote $L \setminus k = (e'_i)_{1 \leqslant i \leqslant n-1}$ where $e'_i = e_i$ if $i < k$, and $e'_i = e_{i+1}$ otherwise. I.e., $L \setminus k$ is the log L in which the kth event has been removed. Under the same premisses, we denote $L_{<k}$ the log $(e_i)_{1 \leqslant i < k}$.

Definition 1. (Evaluation functions) An evaluation function with memory k*, or* k-evaluation function*, is a function* $\varphi : (\mathcal{I} \times \mathcal{P})^k \to \mathbb{B}$.

We say that a log $L = (e_i)_{1 \leqslant i \leqslant n}$ of length n is *accepted* by a k-evaluation function φ if, for all $k \leqslant i \leqslant n$, we have $\varphi(e_{i-(k-1)}, \ldots, e_i) = \top$. Conversely, for each $k \leqslant i \leqslant n$ such that $\varphi(e_{i-(k-1)}, \ldots, e_i) = \bot$, we say that the event i is an *anomaly*.

Let us now define how an evaluation is applied on a log that may contain anomalies.

Definition 2. (Application of a k*-evaluation function) Given a log* L *of length* n *and a* k*-evaluation function* φ*, the* application of φ on L at step $k \leqslant i \leqslant n$ *is denoted* $\mu(\varphi, L, i)$*. It is defined if* φ *accepts* $L_{<i}$ *and when this is the case, we have:*

$$\mu(\varphi, L, i) = \begin{cases} L, & \text{if } i = n + 1; \\ \mu(\varphi, L, i + 1), & \text{if } \varphi(e_{i-(k-1)}, \ldots, e_i) = \top; \\ \mu(\varphi, L \setminus i, i), & \text{if } \varphi(e_{i-(k-1)}, \ldots, e_i) = \bot. \end{cases}$$

The application of φ *on* L *is denoted* $\mu(\varphi, L)$ *and is equal to* $\mu(\varphi, L, 0)$.

We call the result of the application of an evaluation function φ on a log L the φ-accepted subsequence of the elements of a log L. Elements that have been eliminated are said to have been rejected by φ.

The Intrusion detection problem. We let $\mathbb{L} = \{L_i\}_{i\in\mathbb{N}}$ be a set of logs. The *intrusion detection computation problem* consists in computing a parameter k and a k-evaluation function $\varphi_{\mathbb{L}}$ such that, for all $L \in \mathcal{L}$, we have $\mu(\varphi_{\mathbb{L}}, L) \in \mathbb{L}$. Unsurprisingly, given the generality of the notions introduced, we have:

Theorem 1. Every k-evaluation function reckognises a regular language.

Proof (Sketch) Given a k-evaluation function φ, we construct a finite automaton $\mathcal{A}_\varphi$ as follows:

– All states are final, and are the elements in the $\bigcup_{0\leqslant l\leqslant k-1}(\mathcal{I}\times\mathcal{P})^l$;

– Letters are all the elements in $\mathcal{I}\times\mathcal{P}$;

– There is a transition $(e_1,\ldots,e_{k-1}) \to^e (e_2,\ldots,e_{k-1},e)$ if, and only if, $\varphi(e_1,\ldots,e_{k-1},e) = \top$;

– For $l < k-1$, there is a transition $(e_1,\ldots,e_l) \to^e (e_1,\ldots,e_l,e)$;

– The initial state is the state $()$.

It is clear that $\mathcal{A}_\varphi$ accepts a log L if, and only if, $\mathcal{A}_\varphi$ accepts L.

As a corollary of Theorem 1, since sets of logs $\mathbb{L}$ are not assumed to be rational, the intrusion detection problem usually does not have a solution. Beyond this formal impossibility, one can also note that the set of possible k-evaluation functions, even for $k = 1$, is too large to be computed explicitely.

For practical purposes, one thus has to rely on heuristics to find evaluation functions that are of practical use to detect intrusion.

4.2. Criteria for relevant evaluation functions. Since it is unlikely that the possible logs of a non-trivial system form a rational langugage, we aim at learning a *flight envelope* for the system under analysis by overapproximating the set of possible logs of the system with a rational language. Towards this end we try to compute, given the values occurring in the different fields of the legitimate messages, what the possible acceptable values are for these fields. Just as to locate a point in space there is an infinite number of possible basis in which the coordinates of the point can be expressed, there is in principle an infinite number of ways of looking at values of the fields and their interactions one with another.

For this implementation, we have focused on two sources of regularity in the messages normally exchanged on the CAN bus:

– as a car is an example cyber-physical system, some field values represent "*physical*"values, and while their range may encompass the whole

set of possible values, they are likely to change slowly from one message of a given ID to the next;

– the ECU communicating over the CAN bus run computer programs, and those programs are likely to test for the presence of a specific value in the message, or its membership in a small set of possible values. Legitimate messages sent on the bus are constructed so as to pass these tests.

These considerations, explored in more details below, led us to consider testing whether the value of a field stays in a small set, and whether the value of its differential stays in a similarly small state. The set of all possible tests is the *test space*. The tests that are consistently passed by all the messages of a given ID in a log are considered to be characteristic of that message ID, and the tests themselves are the *characteristic functions*.

Methodology. Each log file is read twice. In the first reading, anomalies are removed from the log file, and the analyzer computes for each message ID and each field a subset of the characteristics functions so that:

– that subset is small enough;

– each message occurring in the log pass at least one of the test.

When no small subset is available, the analysis of the field is considered to be inconclusive. During the second read, for each message, the monitor scans each field for which at least one of the value or differential analysis was conclusive. Each field is accepted if one of the retained tests on its value succeeds, and is rejected otherwise. The message is accepted if no field has been rejected.

Methodology on the choice of the test space. The first step consists in choosing a set of simple tests that are likely to be relevant. The space of all possible message tests will then be all the possible conjunctions and disjunctions of these simple tests. We model packets by an ID and a sequence of bytes, *i.e.* 256-valued integers. This ID determines a class to which each packet belongs. We assume that all packets in a given class are similar enough so that some tests exist that are valid on all messages on the class and are not vacuous.

In principle the test space encompasses all boolean functions on messages or sequences of messages. However a succinct analysis already delineates a few types of tests that may be useful for the analysis of logs:

– some tests are related to the syntactic content of the packet, such as the presence of a padding constant or the presence of a specific value, denoting *e.g.* a more precise type for the packet;

– some tests are computed on the whole packet, such as an error-correcting code;

– some tests are domain specific and relate to the possible evolution of physical data between consecutive packets or the set of possible values of some data;

– some tests depend on the internal state of the devices, a packet being acceptable at some point of their execution but not at another point.

For the sake of simplicity we consider in this paper only tests performed independently on the different fields of messages, as well as on their ID. That is, we consider only the first and third cases of the preceding list. We are currently working on implementing the second (whole message tests) and fourth (with an online process mining algorithm).

Automatic fields. A field is *automatic* if the device receiving and accepting this packet tests whether the value of the field is equal to a constant in its program. It is expected that, if different packets can be sent from one device to another, at least one automatic field exists so that the receiver can derive the type of the received packet. The statistical characteristic of such fields are that they should have only *a few* legitimate values, and that these values should have no other detectable relations. However, the difference between these values can be arbitrary as it is simply a case of a few bits switching value.

There is obviously some arbitrariness in deciding what *a few* means. Since the tests performed are not based on any hints from the protocol, we have arbitrarily decided to define a small set of different values to be the square root of the total number of possible different values, that is less than 16 values among the 256 possible ones. Tests relevant to automatic fields are *value tests* in which we record all the different values occurring in a field during training. If the number of different values is more than 16, the analysis is considered to be inconclusive, and no value test is performed on that field for that message ID during monitoring. Otherwise we verify during monitoring that the value in that field for a message is among the ones seen during training. To sum up, value tests are a conjunction, on all fields f, of a disjunction $f = v_1 \vee \ldots \vee f = v_k$ with $k \leqslant 16$, or of the $true$ constant $\top$ if more than 16 different values have been encountered.

Physical values. These are values that are assumed to evolve slowly. For these values we assume a bound on the difference between the value present in the current packet *wrt* the value occurring in the last preceding similar packet. For these fields the analyzer keeps track of the value in the last accepted message and compares that value with the one in the current message. As in the case of value tests, these difference tests are performed during monitoring only if a small (less than 16, again based on a square root consideration) number of changes have been observed during the training phase. Re-using the same notation as above, but now denoting f the value

of a field in the last accepted packet, and f' its value in the packet under analysis, difference tests are a conjunction, on all fields f, of a disjunction $(f' - f) = v_1 \vee \ldots \vee (f' - f) = v_k$ with $k \leqslant 16$, or of the $true$ constant $\top$ if more than 16 different values have been encountered for the difference between the values for that field between a message and its predecessor.

Random values. There are fields for which no relation was found in the data set among the ones that were searching for. In the data sets considered, a post-analysis of the rules has shown that in several cases these fields are often related with the physical value fields, and that the data conveyed were actually 2-bytes values. The analyzer does not perform any test on these fields, as per the construction described both the value and the difference tests are reduced to the $\top$ constant for these.

4.3. Characteristic functions. We sum up the presentation above with the following criteria on a sufficiently good evaluation function φ:

1. We can forget by relations between the content of different message IDs defining a log with multiple IDs as the coproduct of logs each restrict to one single ID ι, *i.e.*, each log L is assumed equal to $\bigoplus_{\iota in \mathcal{I}} \pi_\iota(L)$;

2. The decomposition of each payload into a set of meaningful fields means that each φ_ι can further be decomposed into evaluation functions specific each field f, *i.e.*, $\varphi_\iota = \bigwedge_{f \in \mathcal{F}_\iota} \varphi_{\iota,f}$;

3. To take into account physical values, it suffices to assume that each $\varphi_{\iota,f}$ is a 2-evaluation functions, and that it suffices to consider the difference between the present value of field and its former value;

A final criterion, not introduced above but that we believe is necessary for the stability of the learning phase, is to refrain from having forbiden values, *e.g.* saying that the value in the field 0 will never be 129. As a heuristic these criteria can certainly be relaxed, but we have already obtained good results even though they may seem very restrictive.

In order to define characteristic functions, it suffices now to introduce, for each ID ι, a set of fields $\mathcal{F}_\iota$. Each field $f \in \mathcal{F}_\iota$ is a function $f : \{\iota\} \times [0, 255]^8 \to \mathbb{Z}$. Also for each field $f \in \mathcal{F}_\iota$ we introduce two sets of values $V_{\iota,f}$ and $D_{\iota,f}$ which, according to the above discussion, can either be finite and of cardinal between 1 and 16, or $\mathbb{Z}$.

Definition 3. (Characteristic functions) A characteristic function *for an ID ι with a set of fields $\mathcal{F}_\iota$ is a 2-evaluation function $\varphi_\iota(e, e')$ of the form:*

$$\varphi_\iota(e, e') = \bigwedge_{f \in \mathcal{F}_\iota} \bigvee_{v \in V_f} (f(e') = v) \ \wedge \bigwedge_{f \in \mathcal{F}_\iota} \bigvee_{v \in D_f} (f(e') - f(e) = v).$$

Finally, the log analysis function is the function φ such that $\mu(\varphi, \bigoplus_{\iota in \mathcal{I}} \pi_\iota(L)) = \bigoplus_{\iota in \mathcal{I}} \mu(\varphi_\iota, \pi_\iota(L))$.

Implementation. Characteristic functions, and thus the log analysis functions they define, have been implemented in C. As a first step, the log is translated if necessary into a binary file which is then mapped to an array of structures using `mmap` call, with each structure representing a packet. Records in this array are then analyzed independently by two modules, one tracking for each ID and for each field the number of different values, until the threshold 16 is reached, the other tracking the differences between consecutive values, again for each ID and for each field of that ID. Each analysis module constructs a balanced binary tree mapping an ID and a field to the result of the analysis on this ID for this field. The `monitor` module then uses this structure to parse and iterate over another log file to classify each packet as to whether it should be accepted or not. The complexity of treating each event in this architecture is $\Theta(\log |\mathcal{I}|)$, as we assume the number of fields is bounded, and thus the number of elements in the balanced binary trees is $\Theta(|\mathcal{I}|)$. Thus the treatment time for a log of N events with K different IDs is $\Theta(N \cdot \log K)$, both for learning and monitoring.

Memroy footprint. During training the entire log file is virtually available in memory, and we rely on the operating system to optimize speed and memory consumption. During the rule evaluation the memory needed by the monitor is linear to both the number of different IDs and in the number of fields within the payload. We note however that since each φ_ι is a 2-evaluation function, both the learning and the monitoring can be performed online, with space requirements of $\Theta(\log |\mathcal{I}|)$.

5. Implementation and Evaluation of the Characteristic Functions Method. Our characteristic functions method attempts firstly to classify messages into classes, and secondly to characterize messages in a given class by the set of rules they are required to pass. The `monitor` module only implements tests that are satisfied by all messages in a given class. The classification tool then outputs the specific rules that are to be used in message classification for each individual class. This information is provided in a human-readable format and may potentially be useful in future research as well.

First, it permits to compute the probability that a random message satisfies all the tests in the class, and thus allows us to evaluate the robustness of the monitor against the injection of random messages. Assuming that in a given class there are n fields classified as automatic and m fields classified as physical, and that tests on fields all accept the maximum of 16 values, a random message in that class has a probability $(\frac{16}{256})^{n+m} = 2^{-4\cdot(n+m)}$ to be

Table 2. Intrusion detection results

Scenario	**CF**			**NN**			**NN'**		
Measure	F_1	PPV	TPR	F_1	PPV	TPR	F_1	PPV	TPR
CSA_1	1	1	1	1	1	1	0	0	0
CSA_2	1	1	1	1	1	1	0	0	0
CSA_3	1	1	1	1	1	1	0	0	0
CSA_{1m}	1	1	1	1	1	1	0.999	1	0.999
CSA_{2m}	1	1	1	1	1	1	0.999	1	0.999
CSA_{3m}	1	1	1	1	1	1	0.999	1	0.999
$Fuzz_1$	1	1	1	0.999	0.998	1	0.993	0.994	0.992
$Fuzz_2$	1	1	1	0.996	0.992	1	0.981	0.980	0.983
$Fuzz_3$	1	1	1	0.987	0.975	1	0.974	0.957	0.991
$MECTA$	1	1	1	0	0	0	0	0	0
$MECTA_m$	1	1	1	0	0	0	0	0	0
MSA_1	1	1	1	1	1	1	0	0	0
MSA_2	1	1	1	1	1	1	0	0	0
MSA_3	1	1	1	1	1	1	0	0	0
MSA_{1m}	1	1	1	1	1	1	0.138	0.989	0.074
MSA_{2m}	1	1	1	1	1	1	0.466	0.998	0.304
MSA_{3m}	1	1	1	1	1	1	0.115	0.997	0.061
$RLoff_1$	1	1	1	1	1	1	0	0	0
$RLoff_2$	1	1	1	1	1	1	0.001	1	0.001
$RLoff_3$	1	1	1	1	1	1	0	0	0
$RLoff_{1m}$	1	1	1	1	1	1	0	0	0
$RLoff_{2m}$	1	1	1	1	1	1	0.438	1	0.280
$RLoff_{3m}$	0	1	0	1	1	1	0.556	1	0.385
$RLon_1$	1	1	1	0.990	1	0.980	0	0	0
$RLon_2$	1	1	1	1	1	1	0	0	0
$RLon_3$	1	1	1	1	1	1	0	0	0
$RLon_{1m}$	0.001	1	0	0.987	1	0.975	0.997	0.995	0.999
$RLon_{2m}$	0.340	1	0.204	1	1	1	0.994	0.997	0.991
$RLon_{3m}$	0.578	0.487	0.712	1	1	1	0.998	0.995	0.999

accepted. This small but non-negligible probability explains the occurrences of false negatives in Table 2, where evaluation results for this approach are labelled with `cf` for *characteristic functions*.

Scenario: log-file with simulated attacks; Precision (Positive Predictive Value) $PPV = \frac{TP}{TP+FP}$; Recall (True Positive Rate) $TPR = \frac{TP}{TP+FN}$; F1 Score : $F1 = 2 * \frac{PPV*TPR}{PPV+TPR}$) Different classification scenarios are *characteristic functions* (`cf` or neural networks trained with either the original ORNL intrusion sets (`nn`) or randomly introduced intrusion messages (`nn'`).

Second, given that the rules generated implement simple tests, it is also in theory possible for a human to better understand the system by looking at the rules produces, and eventually produce new (and less generic) tests beyond those described in this paper. A side result of this is that it is also quite easy to build a fake traffic that will be accepted by a monitor once we know its rules.

Third, it permits to focus further classification work on classes for which only a few fields are tested. For example, some poorly classified messages

seem to be frames in a more complex *Multi-Frame Message* (MFM). To handle this case we plan in future works to implement MFM protocol recognition. Also, and though this is outside of the scope of this paper, a manual analysis of the rules produced and of the messages in these classes strongly suggests new test functions, such as counter and checksum detection, to handle these currently poorly handled cases.

In addition to the discussion above, the results of experiments in Table 2 show next to no *false positive* classifications. This is further visualized in Figure 2, where only one column for the characteristic functions method, here marked as `logan`, can be seen. The evaluation results show that though arbitrarily selected, the heuristic threshold of 16 is not too high as it does not classify a field that contains random values into an automatic field, *i.e.* no over-fitting has been observed. This however should not be interpreted as an impossibility for our method to suffer from over-fitting. Especially a training data set which is too short would tend to produce illegitimate value tests, *e.g.* for the fields recording the timestamp of the packet. The high number of *false positives* on one intrusion scenario however, shows a behaviour yet to completely evaluated, where intrusions that alter existing messages instead of only introducing new messages potentially cause the internal state of the *characteristic functions* classifier to reject every message after the first malicious intrusion message has been classified. In a real-life scenario this would potentially not be harmful, due to the fact, that an intrusion has to be detected in order for this to occur.

For a better understanding of the classifier's errors, we visualized the number of FP and FN in a form of bar charts that are presented in Figures 2 and 3, so one can see how errors are distributed over different attack scenarios and classifiers. The `logan` classifier seen said Figures corresponds to the *characteristic functions* approach, whereas `nn_orig` and `nn_fuzzed` correspond to the different training scenarios for the neural network approach, discussed in Section 6.

We also map classification results in form of radial bar charts where blue represent TP and TN, red – FP, and orange – FN. The example is presented in Figure 4. All classification results in form of radial bar charts are available via the link https://guardeec.github.io/ornl_dataset_vis/visualization.html. You can select a data set and classifier type to view the corresponding result. The CAN ID is displayed by clicking on the bar.

As can be seen in Table 2, the results are very encouraging against the different attacks considered. It is to be noted that using knowledge of the results and models from the analysis modules, it would potentially be easy to

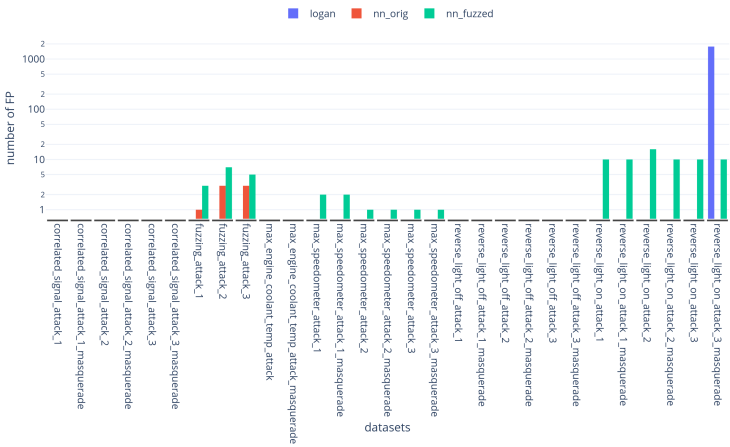


Fig. 2. Classification FP errors for different data sets

construct attacks (*i.e.*, introduction of additional malicious messages on the bus) that follow a pattern that will be accepted by the analyzer.

In addition to the intrusion detection performance we have also evaluated the number of classified messages per second from all test scenarios, which averaged at approx. 1700 messages per second. This test was performed on a Raspberry Pi 3 Model B to test the performance of the classifier on a device similar to what could be used as an edge node in a vehicle.

6. Baseline Benchmark: Artificial Neural Network. As a benchmark for the evaluation of our approach we implemented an artificial neural network approach using the Tensorflow Keras API [30, 31]. Neural networks are the standard for deep learning and can model very complex nonlinear relationships. A fully connected neural network utilizes a number of layers with each layer supporting an arbitrary number of neurons. Data is propagated from the input to the output layer using weighted connections between the neurons of these layers. Specifically a multilayer perceptron (MLP) based on the `Sequential` model from the `keras` Tensorflow package with two hidden layers of 25 neurons each was used. This results in a model with ~ 1200 trainable parameters. We specifically designed the network to perform well on weaker in-vehicle edge devices. As the activation function for the hidden layers we selected *rectified linear unit* (ReLU), which is computationally cheap. In total we trained on the data set for a learning phase of 20 epochs, with a validation split of 0.2, so 20%

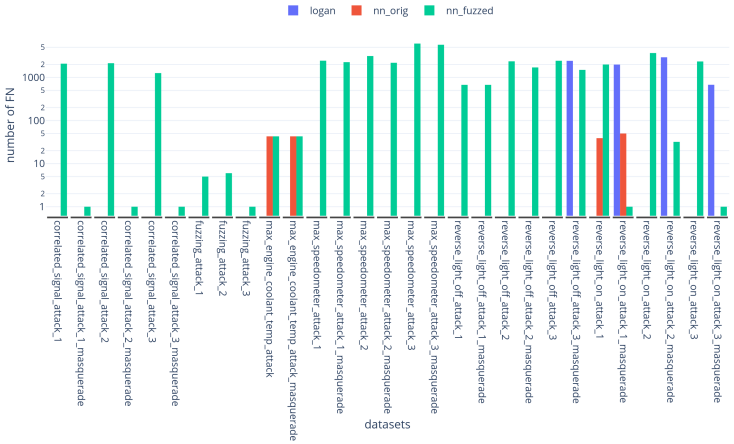


Fig. 3. Classification FN errors for different data sets

of the input data was used for validation. For the *loss* function of the model we decided to use *binary cross-entropy*, which is based on a classification of values between 0 and 1 and is best suited for binary classification, as is required for our training data. In addition to that the *Adam* [32] optimizer was used.

The data logs are then preprocessed into a structure where for each message m^i with arbitration ID i and payload p_{m^i} a input vector (i, p_{m^i}, p_{m-1^i}) with the payload of the previous message with the same ID, is created. With first message of each ID, where there is no previous payload available, a vector of zeros is used as payload. The timing of the message is disregarded in this approach. This structure was selected to make the neural network approach as comparable to the characteristic functions approach possible, by providing the same information for the classifying process as in the case with characteristic functions.

For each of the intrusion scenarios described in Section 3 a separate model was trained, where scenarios that consist of more than one log file are merged into one model. For the training only the non modified log files from the ORNL road data set were used, due to the fact that the masquerade intrusions alter the structure of the log and can potentially impede training.

To improve our evaluation using neural networks we have designed two different evaluation scenarios. One scenario utilizes the original data logs for training data, while the other uses artificially generated intrusion messages,

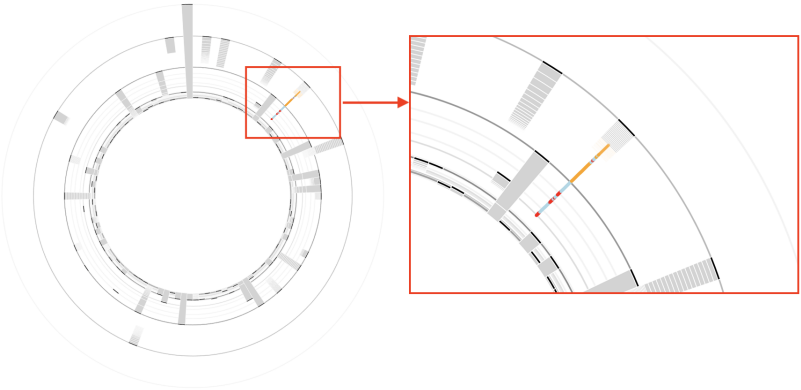


Fig. 4. Classification errors mapped on radial bar chart: blue – TP and TN, red – FP, orange – FN

based on the original intrusions and introduced with a normal distribution over the course of the complete log file. This does not alter the content of the intrusion messages introduced and figuratively represents the case that in real operation the attack is known but the context of the attack is not. Both scenarios have been additionally trained on a version of their respective log files, where all intrusion messages have been deleted in an attempt to improve the classification of normal driving behaviour and minimize the number of false positive classification.

In total six models were trained per evaluation scenario, namely `CSA` (correlated signal attack), `Fuzz` (fuzzing attack), `MECTA` (max engine coolant temp attack), `MSA` (max speedometer attack), `RLoff` (reverse light off attack) and `RLon` (reverse light on attack).

The results are shown in Table 2, where the evaluation scenario using original log files is annotated with *nn* and the scenario using artificially generated messages with *nn'*.

As expected the results for the *nn* scenario are in most scenarios near perfect, except for the `MECTA` intrusion scenario, which contained too few intrusion messages for reliable training. In most other scenarios all introduced intrusion messages were classified correctly as intrusions, as indicated by a value of 1 in the TPR_{nn} column. For the fuzzing attacks a below a 1 value in the PPV_{nn} column indicates the occurrence of *false positive* values in classification, a close observation here shows misclassification here happens mostly at the beginning of the log, where a zero value vector was used in the input vector for the message, as described above.

The results for the training with artificially introduced intrusion data, which can shown in the *nn'* column of Table 2, are more diverse. The classifier models have shown that often for intrusion scenarios, where additional messages were introduced to obfuscate the normal behaviour, the classification performance of the models trained on artificially placed intrusion is zero or close to zero. This shows that the context of the messages, most importantly the previous message is decisive for correct classification. For the masquerade scenarios of each intrusion, many of the introduced and modified messages were classified correctly. A high *positive predictive value* here shows that the number of false positive classifications is close to zero, whereas the *true positive rate* varies significantly with different log files. These scenarios show, that despite the low classification rates on the non-masquerade versions of the logs, the models are able to detect derivations from normal behaviour in the log files.

The results here highlight the complexity of the intrusion scenarios from the ORNL Road data set. The *nn'* model evaluation signifies that even if the message structure of an intrusion scenario is known, the correct classification is non-trivial.

To provide a better performance comparison to the CF approach not only in terms of classification accuracy, but also in regards to time performance we have also run all evaluations on a Raspberry Pi 3 Model B for the ANN classifier. On this relatively low-performance device the ANN was only able to evaluate an average of 150 messages per second, which is less than a tenth of the performance shown by the CF classifier.

7. Conclusion. We have seen in previous work [6] that artificial neural network approaches to anomaly detection deliver good results but that it is hard to implement this kind of detection in-vehicle because of restrictions with respect to on-board resources of typical ECUs used in vehicular systems. Thus, we have started to analyze logs using a bind and branch approach that was very accurate but lacked robustness. From this experience we built a log analyzer in C that focused on payload bytes having either a small set of different values or a small set of possible changes. We have evaluated this characteristic functions approach on state-of-the-art CAN bus intrusion data from real-life intrusion scenarios and obtained results that are significantly more robust and accurate in comparison to a standard implementation of an artificial neural network classifier. The evaluations regarding the time performance of both approaches have also shown a significant margin between both approaches with characteristic functions being able to evaluate ten times more messages with the same time compared to even relatively small artificial neural network.

The approach has shown to classify small derivation from standard behaviour in log files well without the occurrence of *false positive* classifications, which is an important trait for the integration in a real automotive environment. As an extension of our work in [4], we have shown here that even timing opaque attacks from the sophisticated ORNL data set which do not disrupt normal timing nor CAN ID distributions can be found with our method.

We will work in the near future on refining the analysis to *guess* the functions employed by the devices to test whether the packet shall be accepted. We plan to extend our approach to CAN with flexible data-rate (CAN-FD) which is an extension of the original CAN bus protocol with higher bandwidth. Furthermore, we work on a hybrid method where artificial neural networks are used offline to improve the rules of a rule-based in-vehicle IDS.

References

1. Müller-Quade J., Backes M., Buxmann P., Eckert C., Holz T. *et al.* Cybersecurity research: Challenges and course of action. *Tech. rep., Karlsruher Institut für Technologie (KIT).* 2019.
2. Miller C., Valasek C. Remote exploitation of an unaltered passenger vehicle. *Tech. rep., IOActive Labs.* August 2015.
3. UN Regulation No. 155 [Uniform provisions concerning the approval of vehicles with regards to cyber security and cyber security management system]. Available at: www.eur-lex.europa.eu/eli/reg/2021/387/ojOnline. (accessed 29-Apr-2021).
4. Chevalier Y., Rieke R., Fenzl F., Chechulin A., Kotenko I. Ecu-secure: Characteristic functions for in-vehicle intrusion detection. Proceedings of the International Symposium on Intelligent and Distributed Computing. 2019. pp. 495–504.
5. Hacking and Countermeasure Research Lab (HCRL). [Car-Hacking Dataset for the intrusion detection]. Available at: http://ocslab.hksecurity.net/Datasets/CAN-intrusion-dataset. (accessed 28-Jun-2018).
6. Berger I., Rieke R., Kolomeets M., Chechulin A., Kotenko I. Comparative study of machine learning methods for in-vehicle intrusion detection. Proceedings of the ESORICS 2018 International Workshops, CyberICPS 2018 and SECPRE 2018, Barcelona, Spain, September 6-7, 2018, Revised Selected Papers. 2019. vol. 11387. pp. 85–101.
7. Verma M., Iannacone M., Bridges R., Hollifield S., Kay B. Combs F. Road: The real ornl automotive dynamometer controller area network intrusion detection dataset (with a comprehensive can ids dataset survey & guide. ArXiv preprint arXiv:2012.14600. 2020.
8. Studnia I., Nicomette V., Alata E., Deswarte Y., Kaâniche M., Laarouchi Y. Security of embedded automotive networks: state of the art and a research proposal. Proceedings of the SAFECOMP 2013 - Workshop CARS of the 32nd International Conference on Computer Safety, Reliability and Security. 2013.
9. Wolf M., Weimerskirch A., Paar C. Security in Automotive Bus Systems. Proceedings of the Workshop on Embedded Security in Cars. 2014. pp. 1–13.
10. ENISA Cyber security and resilience of smart cars. *Tech. rep., ENISA*. 2016.
11. Metzker E. Reliably detecting and defending against attacks. Available at: https://assets.vector.com/cms/content/know-how/_technical-articles/Security_Intrusion_Detection_AutomobilElektronik_202003 _PressArticle_EN.pdf. (accessed 28-Apr-2021).

12. Choi W., Joo K., Jo H., Park M., Lee D. Voltageids: Low-level communication characteristics for automotive intrusion detection system. *IEEE Transactions on Information Forensics and Security.* 2018. vol. 13. pp. 2114–2129.
13. Cho K., Shin K. Fingerprinting electronic control units for vehicle intrusion detection. Proceedings of the 25th USENIX Security Symposium, USENIX Security 16, Austin, TX, USA, August 10-12, 2016. 2016. pp. 911–927.
14. Larson U., Nilsson D., Jonsson E. An approach to specification-based attack detection for in-vehicle networks. Proceedings of the Intelligent Vehicles Symposium, 2008 IEEE. 2008. pp. 220–225.
15. Hoppe T., Kiltz S., Dittmann J. Security threats to automotive CAN networks – practical examples and selected short-term countermeasures. *Reliability Engineering & System Safety.* 2011. vol. 96. pp. 235–248.
16. Müter M., Asaj N. Entropy-based anomaly detection for in-vehicle networks. Proceedings of the 2011 IEEE Intelligent Vehicles Symposium (IV). 2011. pp. 1110–1115.
17. Studnia I., Alata E., Nicomette V., Kaâniche M., Laarouchi Y. A language-based intrusion detection approach for automotive embedded networks. Proceedings of the 21st IEEE Pacific Rim International Symposium on Dependable Computing (PRDC 2015). 2014. pp. 1–12.
18. Song H., Kim H., Kim H. Intrusion detection system based on the analysis of time intervals of CAN messages for in-vehicle network. Proceedings of the 2016 international conference on information networking (ICOIN). 2016. vol. 3. pp. 63–68.
19. Wei Z., Yang Y., Rehana Y., Wu Y., Weng J., Deng R. IoVShield: An Efficient Vehicular Intrusion Detection System for Self-driving. Proceedings of the International Conference on Information Security Practice and Experience. 2017. pp. 638–647.
20. Rieke R., Seidemann M., Talla E., Zelle D., Seeger B. Behavior analysis for safety and security in automotive systems. Proceedings of the Parallel, Distributed and Network-Based Processing (PDP), IEEE Computer Society. 2017. pp. 381–385.
21. Levi M., Allouche Y., Kontorovich A. Advanced analytics for connected cars cyber security. Proceedings of the 87th Vehicular Technology Conference (VTC Spring), IEEE. 2017. vol. abs/1711.01939.
22. Narayanan S., Mittal S., Joshi A. Obd securealert: An anomaly detection system for vehicles. Proceedings of the IEEE Workshop on Smart Service Systems (SmartSys 2016). 2016. pp. 1–7.
23. Theissler A. Anomaly detection in recordings from in-vehicle networks. Proceedings of Big Data Applications and Principles First International Workshop, BIGDAP 2014. 2014. vol. 23. P. 26.
24. Kang M., Kang J. A novel intrusion detection method using deep neural network for in-vehicle network security. Proceedings of the 83rd Vehicular Technology Conference (VTC Spring), IEEE. 2016. pp. 1–5.
25. Marchetti M., Stabili D. Anomaly detection of CAN bus messages through analysis of id sequences. Proceedings of the 2017 IEEE Intelligent Vehicles Symposium (IV). 2017. pp. 1577–1583.
26. Chockalingam V., Larson I., Lin D., Nofzinger S. Detecting attacks on the CAN protocol with machine learning. *Annu EECS.* 2016. vol. 558. no.7.
27. Taylor A., Leblanc S., Japkowicz N. Probing the limits of anomaly detectors for automobiles with a cyber attack framework. *IEEE Intelligent Systems.* 2018. vol. 33. no. 2. pp. 54–62.
28. Al-Jarrah O., Maple C., Dianati M., Oxtoby D., Mouzakitis A. Intrusion detection systems for intra-vehicle networks: A review. *IEEE Access.* 2019. vol. 7. pp. 21266–21289.

29. Kolomeets M., Chechulin A., Kotenko I. Visual analysis of CAN bus traffic injection using radial bar charts. Proceedings of the 1st IEEE International Conference on Industrial Cyber-Physical Systems (ICPS-2018). 2018. pp. 841–846.
30. Abadi M., Barham P., Chen J., Chen Z., Davis A., Dean J., Devin M., Ghemawat S., Irving G., Isard M., et al. Tensorflow: A system for large-scale machine learning. Proceedings of the 12th USENIX symposium on operating systems design and implementation (OSDI 16). 2016. pp. 265–283.
31. Chollet F. Keras. Available at: https://github.com/fchollet/keras. (accessed 28-Apr-2021).
32. Kingma D., Ba J. Adam: A method for stochastic optimization. ArXiv preprint arXiv:1412.6980. 2014.

Chevalier Yannick — Dr. Hab., Assistant Professor, AI department, Université de Toulouse, IRIT. Research interests: formal methods for security, verification of cryptographic protocols. The number of peer reviewed publications — 46. ychevali@irit.fr; www.irit.fr/ Yannick.Chevalier/; 118, route de Narbonne, 31062, Toulouse, Cedex 7, France; office phone: +33-561556091.

Fenzl Florian — Researcher, Department Cyber-Physical Systems Security, Fraunhofer Institute for Secure Information Technology SIT. Research interests: automotive security, artificial intelligence, machine learning for anomaly detection. The number of peer reviewed publications — 2. florian.fenzl@sit.fraunhofer.de; Rheinstrasse 75, 64295 Darmstadt, Germany; office phone: +49-6151869116.

Kolomeets Maxim — PhD Student, Juniour Researcher, Laboratory of Computer Security Problems, St. Petersburg Federal Research Center of the Russian Academy of Sciences. Research interests: security data visualisation, social networks security. The number of peer reviewed publications — 21. kolomeec@comsec.spb.ru; 14th line of V.O. 39, 199178, Saint-Petersburg, Russia; office phone: +7(812)328-7181.

Rieke Roland — Dr. rer. nat., Senior Scientist, Department Cyber-Physical Systems Security, Fraunhofer Institute for Secure Information Technology SIT. Research interests: design principles for secure, scalable systems and model-based predictive security analysis. The number of peer reviewed publications — 46. roland.rieke@sit.fraunhofer.de; Rheinstrasse 75, 64295 Darmstadt, Germany; office phone: +49-6151869116.

Chechulin Andrey — PhD, Leading Researcher, Laboratory of Computer Security Problems, St. Petersburg Federal Research Center of the Russian Academy of Sciences. Research interests: computer network security, intrusion detection, analysis of vulnerability, security visualization, embedded systems security. The number of peer reviewed publications — 80. chechulin@comsec.spb.ru; 14th line of V.O. 39, 199178, Saint-Petersburg, Russia; office phone: +7(812)328-7181.

Krauss, Christoph — Dr., Professor, Head, Department Cyber-Physical Systems Security, Fraunhofer Institute for Secure Information Technology SIT. Darmstadt University of Applied Sciences. Research interests: automotive security and privacy, railway security, intelligent energy networks security, trusted computing, network security, efficient and post quantum cryptography. christoph.krauss@sit.fraunhofer.de; Rheinstrasse 75, 64295 Darmstadt, Germany; office phone: +49-6151869116.

Acknowledgements. This research is supported by the German Federal Ministry of Education and Research (BMBF) and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE and by the BMBF project VITAF (ID 16KIS0835). Additionally, the project leading to this application has received funding from the European Union's Horizon 2020 research, innovation programme under grant agreement No 883135 and by the budget project 0073-2019-0002.

УДК 004.056.5 **DOI** 10.15622/ia.20.4.4

Я. Шевалье, Ф. Фенцль, М.В. Коломеец, Р. Рике, А.А. Чечулин, К. Краус

ОБНАРУЖЕНИЕ КИБЕРАТАК В ТРАНСПОРТНЫХ СРЕДСТВАХ С ИСПОЛЬЗОВАНИЕМ ХАРАКТЕРИЗУЮЩИХ ФУНКЦИЙ, ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ И ВИЗУАЛЬНОГО АНАЛИЗА

Шевалье Я., Фенцль Ф., Коломеец М.В., Рике Р., Чечулин А.А., Краус. К. **Обнаружение кибератак в транспортных средствах с использованием характеризующих функций, искусственных нейронных сетей и визуального анализа.**

Аннотация. Возможность подключения автономных транспортных средств к сетям порождает новые возможности для атак и, следовательно, потребность в развитии методов кибербезопасности. Таким образом, важно обеспечить, чтобы мониторинг сети в транспортном средстве включал в себя возможность точно обнаруживать вторжение и анализировать кибератаки на основе данных о транспортных средствах и журналов событий транспортных средств с учетом их конфиденциальности. В статье предложен и оценен метод, использующий характеризующую функцию и проведено его сравнение с подходом, основанным на искусственных нейронных сетей. Визуальный анализ соответствующих потоков событий дополняет оценку. Несмотря на то, что метод с характеризующей функцией на порядок быстрее, точность полученных результатов, по крайней мере, сравнима с таковой, полученной с помощью искусственной нейронной сети. Таким образом, этот метод представляет собой перспективный вариант для реализации во встраиваемых системах автомобиля. Кроме того, важным аспектом использования методов анализа в рамках кибербезопасности является объяснимость результатов обнаружения.

Ключевые слова: бзеопасность сети контроллера, обнаружение вторжений, обнаружение аномалий, машинное обучение, автомобильная безопасность, мониторинг безопасности.

Шевалье Янник — Dr. Hab., доцент, кафедра искусственного интеллекта, университет Тулузы, IRIT. Область научных интересов: формальные методы безопасности, верификация криптографических протоколов. Число научных публикаций — 46. ychevali@irit.fr; www.irit.fr/Yannick.Chevalier; 118, route de Narbonne, 31062, Тулуза, Cedex 7, Франция; р.т.: +33-561556091.

Фенцль Флориан — научный сотрудник, департамент безопасности киберфизических систем, институт безопасных информационных технологий им. Фраунгофера (SIT). Область научных интересов: автомобильная безопасность, искусственный интеллект, машинное обучение, обнаружения аномалий. Число научных публикаций — 2. florian.fenzl@sit.fraunhofer.de; Rheinstrasse 75, 64295, Дармштадт, Германия; р.т.: +49-6151869116.

Коломеец Максим Вадимович — аспирант, младший научный сотрудник, лаборатория проблем компьютерной безопасности, Федеральное государственное бюджетное учреждение науки "Санкт-Петербургский Федеральный исследовательский центр Российской академии наук". Область научных интересов: визуализация данных безопасности, безопасность социальных сетей. Число научных публикаций — 21. kolomeec@comsec.spb.ru; 14-я линия В.О. 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-7181.

Рике Роланд — Dr. rer. nat., старший научный сотрудник, департамент безопасности киберфизических систем, институт безопасных информационных технологий им. Фраунгофера (SIT). Область научных интересов: принципы проектирования безопасных масштабируемых систем и прогнозный анализ безопасности на основе моделей. Число научных публикаций — 46. roland.rieke@sit.fraunhofer.de; Rheinstrasse 75, 64295, Дармштадт, Германия; р.т.: +49-6151869116.

Чечулин Андрей Алексеевич — кандидат технических наук, доцент, ведущий научный сотрудник, лаборатория проблем компьютерной безопасности, Федеральное государственное бюджетное учреждение науки "Санкт-Петербургский Федеральный исследовательский центр Российской академии наук". Область научных интересов: безопасность компьютерных сетей, обнаружение вторжений, анализ уязвимостей. визуализация данных безопасности, защита встроенных устройств. Число научных публикаций — 80. chechulin@comsec.spb.ru; 14-я линия В.О. 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-7181.

Краус Кристоф — Ph.D., профессор, руководитель, отдел безопасности киберфизических систем, институт безопасных информационных технологий им. Фраунгофера (SIT). Дармштадтский университет прикладных наук. Область научных интересов: автомобильная безопасность и конфиденциальность, безопасность железных дорог, безопасность интеллектуальных энергетических сетей, надежные вычисления, сетевая безопасность, эффективная и постквантовая криптография. christoph.krauss@sit.fraunhofer.de; Rheinstrasse 75, 64295 Дармштадт, Германия; р.т.: +49-6151869116.

Поддержка исследований. Это исследование поддержано Федеральным министерством образования и исследований Германии (BMBF) и Государственным министерством высшего образования, исследований и искусств земли Гессен в рамках совместной поддержки Национального исследовательского центра прикладной кибербезопасности ATHENE и проекта BMBF VITAF (ID 16KIS0835). Кроме того, проект получил финансирование в рамках исследовательской программы Европейского Союза Horizon 2020, инновационной программы в рамках грантового соглашения № 883135 и бюджетного проекта 0073-2019-0002.

Литература

1. *Müller-Quade J., Backes M., Buxmann P., Eckert C., Holz T. et al.* Cybersecurity research: Challenges and course of action // Tech. rep., Karlsruher Institut für Technologie (KIT). 2019.
2. *Miller C., Valasek C.* Remote exploitation of an unaltered passenger vehicle // Tech. rep., IOActive Labs. August 2015.
3. UN Regulation No. 155 [Uniform provisions concerning the approval of vehicles with regards to cyber security and cyber security management system]. URL: www.eur-lex.europa.eu/eli/reg/2021/387/ojOnline. (дата обращения: 29.04.2021).
4. *Chevalier Y., Rieke R., Fenzl F., Chechulin A., Kotenko I.* Ecu-secure: Characteristic functions for in-vehicle intrusion detection // Proceedings of the International Symposium on Intelligent and Distributed Computing. 2019. pp. 495–504.
5. Hacking and Countermeasure Research Lab (HCRL). [Car-Hacking Dataset for the intrusion detection]. URL: http://ocslab.hksecurity.net/Datasets/CAN-intrusion-dataset. (дата обращения: 28.05.2021).
6. *Berger I., Rieke R., Kolomeets M., Chechulin A., Kotenko I.* Comparative study of machine learning methods for in-vehicle intrusion detection // Proceedings of the ESORICS 2018 International Workshops, CyberICPS 2018 and SECPRE 2018, Barcelona, Spain, September 6-7, 2018, Revised Selected Papers. 2019. vol. 11387. pp. 85–101.
7. *Verma M., Iannacone M., Bridges R., Hollifield S., Kay B. Combs F.* Road: The real ornl automotive dynamometer controller area network intrusion detection dataset (with

a comprehensive can ids dataset survey & guide // ArXiv preprint arXiv:2012.14600. 2020.

8. *Studnia I., Nicomette V., Alata E., Deswarte Y., Kaâniche M., Laarouchi Y.* Security of embedded automotive networks: state of the art and a research proposal // Proceedings of the SAFECOMP 2013 - Workshop CARS of the 32nd International Conference on Computer Safety, Reliability and Security. 2013.
9. *Wolf M., Weimerskirch A., Paar C.* Security in Automotive Bus Systems // Proceedings of the Workshop on Embedded Security in Cars. 2014. pp. 1–13.
10. ENISA Cyber security and resilience of smart cars // Tech. rep., ENISA. 2016.
11. *Metzker E.* Reliably detecting and defending against attacks. URL: https://assets.vector.com/cms/content/know-how/_technical-articles/Security_Intrusion_Detection_AutomobilElektronik_202003 _PressArticle_EN.pdf. (дата обращения: 28.05.2021).
12. *Choi W., Joo K., Jo H., Park M., Lee D.* Voltageids: Low-level communication characteristics for automotive intrusion detection system // IEEE Transactions on Information Forensics and Security. 2018. vol. 13. pp. 2114–2129.
13. *Cho K., Shin K.* Fingerprinting electronic control units for vehicle intrusion detection // Proceedings of the 25th USENIX Security Symposium, USENIX Security 16, Austin, TX, USA, August 10-12, 2016. 2016. pp. 911–927.
14. *Larson U., Nilsson D., Jonsson E.* An approach to specification-based attack detection for in-vehicle networks // Proceedings of the Intelligent Vehicles Symposium, 2008 IEEE. 2008. pp. 220–225.
15. *Hoppe T., Kiltz S., Dittmann J.* Security threats to automotive CAN networks – practical examples and selected short-term countermeasures // Reliability Engineering & System Safety. 2011. vol. 96. pp. 235–248.
16. *Müter M., Asaj N.* Entropy-based anomaly detection for in-vehicle networks // Proceedings of the 2011 IEEE Intelligent Vehicles Symposium (IV). 2011. pp. 1110–1115.
17. *Studnia I., Alata E., Nicomette V., Kaâniche M., Laarouchi Y.* A language-based intrusion detection approach for automotive embedded networks // Proceedings of the 21st IEEE Pacific Rim International Symposium on Dependable Computing (PRDC 2015). 2014. pp. 1–12.
18. *Song H., Kim H., Kim H.* Intrusion detection system based on the analysis of time intervals of CAN messages for in-vehicle network // Proceedings of the 2016 international conference on information networking (ICOIN). 2016. vol. 3. pp. 63–68.
19. *Wei Z., Yang Y., Rehana Y., Wu Y., Weng J., Deng R.* IoVShield: An Efficient Vehicular Intrusion Detection System for Self-driving // Proceedings of the International Conference on Information Security Practice and Experience. 2017. pp. 638–647.
20. *Rieke R., Seidemann M., Talla E., Zelle D., Seeger B.* Behavior analysis for safety and security in automotive systems // Proceedings of the Parallel, Distributed and Network-Based Processing (PDP), IEEE Computer Society. 2017. pp. 381–385.
21. *Levi M., Allouche Y., Kontorovich A.* Advanced analytics for connected cars cyber security // Proceedings of the 87th Vehicular Technology Conference (VTC Spring), IEEE. 2017. vol. abs/1711.01939.
22. *Narayanan S., Mittal S., Joshi A.* Obd securealert: An anomaly detection system for vehicles // Proceedings of the IEEE Workshop on Smart Service Systems (SmartSys 2016). 2016. pp. 1–7.
23. *Theissler A.* Anomaly detection in recordings from in-vehicle networks // Proceedings of Big Data Applications and Principles First International Workshop, BIGDAP 2014. 2014. vol. 23. P. 26.

24. *Kang M., Kang J.* A novel intrusion detection method using deep neural network for in-vehicle network security // Proceedings of the 83rd Vehicular Technology Conference (VTC Spring), IEEE. 2016. pp. 1–5.
25. *Marchetti M., Stabili D.* Anomaly detection of CAN bus messages through analysis of id sequences // Proceedings of the 2017 IEEE Intelligent Vehicles Symposium (IV). 2017. pp. 1577–1583.
26. *Chockalingam V., Larson I., Lin D., Nofzinger S.* Detecting attacks on the CAN protocol with machine learning // Annu EECS. 2016. vol. 558. no.7.
27. *Taylor A., Leblanc S., Japkowicz N.* Probing the limits of anomaly detectors for automobiles with a cyber attack framework // IEEE Intelligent Systems. 2018. vol. 33. no. 2. pp. 54–62.
28. *Al-Jarrah O., Maple C., Dianati M., Oxtoby D., Mouzakitis A.* Intrusion detection systems for intra-vehicle networks: A review // IEEE Access. 2019. vol. 7. pp. 21266–21289.
29. *Kolomeets M., Chechulin A., Kotenko I.* Visual analysis of CAN bus traffic injection using radial bar charts // Proceedings of the 1st IEEE International Conference on Industrial Cyber-Physical Systems (ICPS-2018). 2018. pp. 841–846.
30. *Abadi M., Barham P., Chen J., Chen Z., Davis A., Dean J., Devin M., Ghemawat S., Irving G., Isard M., et al.* Tensorflow: A system for large-scale machine learning // Proceedings of the 12th USENIX symposium on operating systems design and implementation (OSDI 16). 2016. pp. 265–283.
31. *Chollet F.* Keras. URL: https://github.com/fchollet/keras. (дата обращения: 28.04.2021)
32. *Kingma D., Ba J.* Adam: A method for stochastic optimization // ArXiv preprint arXiv:1412.6980. 2014.

УДК 004.896 **DOI** 10.15622/ia.20.4.5

Ф.В. Краснов, И.С. Смазневич, Е.Н. Баскакова

ОПТИМИЗАЦИОННЫЙ ПОДХОД К ВЫБОРУ МЕТОДОВ ОБНАРУЖЕНИЯ АНОМАЛИЙ В ОДНОРОДНЫХ ТЕКСТОВЫХ КОЛЛЕКЦИЯХ

Краснов Ф.В., Смазневич И.С., Баскакова Е.Н. **Оптимизационный подход к выбору методов обнаружения аномалий в однородных текстовых коллекциях.**

Аннотация. Рассматривается задача обнаружения аномальных документов в текстовых коллекциях. Существующие методы выявления аномалий не универсальны и не показывают стабильный результат на разных наборах данных. Точность результатов зависит от выбора параметров на каждом из шагов алгоритма, и для разных коллекций оптимальны различные наборы параметров. Не все из существующих алгоритмов обнаружения аномалий эффективно работают с текстовыми данными, векторное представление которых характеризуется большой размерностью при сильной разреженности.

Задача поиска аномалий рассматривается в следующей постановке: требуется проверить новый документ, загружаемый в прикладную интеллектуальную информационную систему (ПИИС), на соответствие хранящейся в ней однородной коллекции документов. В ПИИС, обрабатывающих юридически значимые документы, на методы обнаружения аномалий накладываются следующие ограничения: высокая точность, вычислительная эффективность, воспроизводимость результатов, а также объяснимость решения. Исследуются методы, удовлетворяющие этим условиям.

В работе изучается возможность оценки текстовых документов по шкале аномальности путем внедрения в коллекцию заведомо инородного документа. Предложена стратегия обнаружения в документе новизны по отношению к коллекции, предполагающая обоснованный подбор методов и параметров. Показано, как на точность решения влияет выбор вариантов векторизации, принципов токенизации, методов снижения размерности и параметров алгоритмов поиска аномалий.

Эксперимент проведен на двух однородных коллекциях нормативно-технических документов: стандартов в отношении информационных технологий и в сфере железных дорог. Использовались подходы: вычисление индекса аномальности как расстояния Хеллингера между распределениями близости документов к центру коллекции и к инородному документу; оптимизация алгоритмов поиска аномалий в зависимости от методов векторизации и снижения размерности. Векторное пространство строилось с помощью преобразования TF-IDF и тематического моделирования ARTM. Тестировались алгоритмы Isolation Forest (изолирующий лес), Local Outlier Factor (локальный фактор выброса), One-Class SVM (вариант метода опорных векторов).

Эксперимент подтвердил эффективность предложенной оптимизационной стратегии для определения подходящего метода обнаружения аномалий для заданной текстовой коллекции. При поиске аномалии в рамках тематической кластеризации юридически значимых документов эффективен метод изолирующего леса. При векторизации документов по TF-IDF целесообразно подобрать оптимальные параметры словаря и использовать метод опорных векторов с соответствующей функцией преобразования признакового пространства.

Ключевые слова: выявление аномалий, выявление новизны, выявление выбросов, однородные текстовые коллекции, уменьшение размерности разреженных пространств, тематическое моделирование.

1. Введение. Алгоритм выявления аномалий в текстовых данных может выступать в качестве одного из компонентов решения во многих прикладных задачах. В полнотекстовом поиске при ранжировании поисковой выдачи можно учитывать количество новой информации, содержащейся в найденных документах. При распределении входящих обращений по отделам или экспертам полезно выявлять заявки, которые не могут быть никем обработаны - например, потому что отправлены как спам или по ошибке. В процессе тематической рубрикации массива документов необходимо помечать документы, которые не вписываются ни в одну из существующих тем. Кроме того, выявление аномалий в коллекциях важно для поддержания их однородности, влияющей на качество решения при выделении фактов, поиске противоречий и в других задачах интеллектуального анализа текста.

Для прикладных интеллектуальных информационных систем (ПИИС) необходимы точные методы обнаружения аномалий, показывающие высокую эффективность на текстовых документах, которые представляются пространством признаков высокой размерности и объединяются в коллекции больших размеров.

Аномалии в текстовых данных требуется обнаруживать на разных уровнях, в зависимости от практической задачи: на уровне тем [1], на уровне документов [2], на уровне предложений [3] и в некоторых случаях на уровне слов (например, в работе [4] как результат поиска аномалий выявляются новые значения слов). Есть и обратная задача - идентифицировать в потоке текстовых данных те элементы (документы, обращения, сообщения), которые не добавляют новой информации к уже накопленной по данной теме [5].

Рассмотрим формальное определение несоответствия нового документа документам исходной коллекции. Допустим, что коллекция обладает распределением P по признакам p_i. Тогда новый документ d может принадлежать либо P, либо другому распределению Q, и во втором случае он является аномальным (отличается новизной). При этом о степени аномальности документа говорит величина $D(P,Q)$, характеризующая расхождение между распределениями P и Q.

В случае с текстами для каждого документа коллекции создается векторное представление в пространстве $p \in \mathrm{R}^{1\times|W|}$, где $|W|$ — натуральное число. Частным случаем является значение $|W|$, равное количеству уникальных слов в коллекции, однако $|W|$ может быть меньше – если в результате снижения размерности получаются плотные вектора, или

больше – когда в качестве термов рассматриваются n-граммы. Для коллекции из $|D|$ документов создается векторное пространство $R^{|D|\times|W|}$. Для добавляемого документа d также может быть получено векторное представление – в пространстве $p \in R^{1\times|\hat{W}|}$. Добавление нового документа в коллекцию может по-разному повлиять на изменение пространства коллекции в зависимости от того, как строится вектор нового элемента – обучая модель заново (полностью обновляя матрицу на основе изменившегося словаря коллекции) либо достраивая модель (формируя вектор нового документа на основе исходного словаря). В первом случае словарь дополняется: $|W| \to |W \cup \hat{W}|$; во втором остается без изменений: $|W| \to |W|$.

Рассмотрим подробнее, что подразумевается под обнаружением аномалии. Результатом работы алгоритма по обнаружению аномалий (неважно, новизны или выбросов) является вектор длины $|D|$, состоящий из рациональных чисел, характеризующих степень аномальности каждого документа. В описаниях алгоритмов эти рациональные числа называются по-разному: степенью или индексом аномальности [6], скорингом либо фактором аномальности (преимущественно в англоязычных работах) [7] или decision-функцией [8]. Аномальными считаются элементы, индекс аномальности которых превышает заданное пороговое значение (или оказывается ниже его), либо к таковым относится некоторое число элементов с наибольшим (или наименьшим) индексом аномальности.

Обнаружение аномалий можно условно разделить на две задачи: обнаружение выбросов (outlier detection) и обнаружение новизны (novelty detection). Обнаружение выбросов подразумевает анализ коллекции с целью выявления нестандартных документов среди всех имеющихся. Задача обнаружения новизны предполагает оценку нового документа и является актуальной для однородной коллекции, собранной по одному или нескольким критериям: предметная область, жанр и стиль текста, назначение информации и способ ее применения, сходная структура документов. Примером такой коллекции может быть база технических заданий, подборка нормативно-правовых актов по определенной сфере деятельности, договоры или доверенности организации. В этом случае задача обнаружения новизны состоит в следующем: если поступающий новый документ «не соответствует» данной коллекции, он должен быть идентифицирован как инородный элемент по отношению к ней. Однородность самой коллекции имеет большое значение для поиска аномально новых документов, поскольку при наличии выбросов

в коллекции значение индекса аномальности у этих элементов может оказаться выше, чем у инородного нового документа.

Данная работа посвящена задаче обнаружения новизны – выявления нового текстового документа как инородного по отношению к заданной коллекции. На рисунке 1 схематично показан каркас исследования.

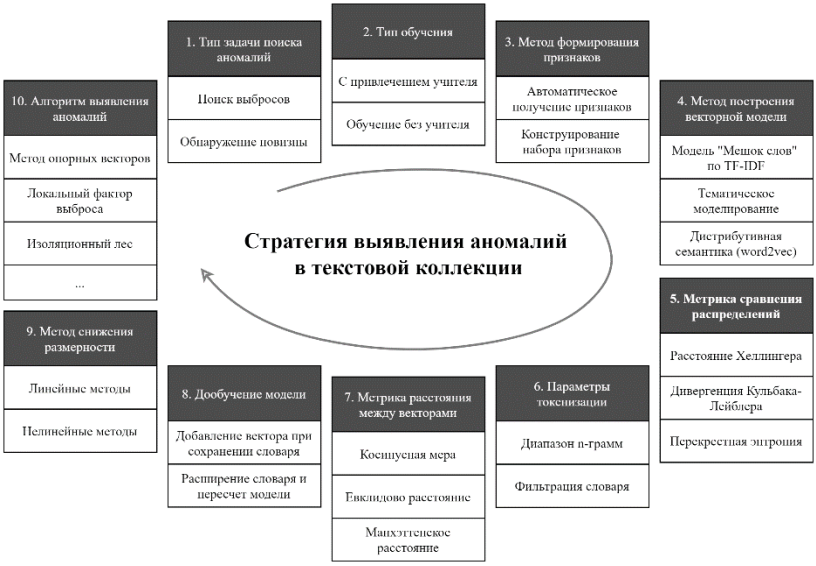


Рис. 1. Каркас исследования: этапы решения задачи обнаружения аномалий в текстовой коллекции

Решение задачи по обнаружению аномалий состоит из нескольких этапов, каждый из которых допускает несколько вариантов подходов или методов, от выбора которых зависит итоговая эффективность всего алгоритма. Нумерация этапов в общем случае соответствует последовательности шагов при поиске аномалий в текстовой коллекции. Однако при выборе определенных методов некоторые этапы могут быть пропущены, тогда как другие становятся обязательными.

Цель настоящего исследования - определить стратегию выбора оптимальных методов и их параметров для обнаружения аномальности нового документа по отношению к существующей текстовой коллекции с учетом ограничений, накладываемых прикладными информационными системами.

С учетом каркаса исследования в качестве методики в данной работе используется оптимизационный подход: определение функционала в пространстве свободных параметров различных стратегий (цепочек алгоритмов) и поиск оптимальной стратегии.

Новизна подхода состоит в предложенной стратегии оптимизации: методы и параметры подбираются для заданной коллекции документов, и при оптимизации используется образец аномальности – заведомо инородный для коллекции документ. Параметры алгоритма считаются оптимальными, когда минимизировано число элементов коллекции, обладающих степенью аномальности выше, чем у инородного документа; в случае однородной коллекции при оптимальных параметрах алгоритма инородный документ обладает максимальной степенью аномальности среди всех элементов коллекции.

Статья состоит из введения, обзора основных алгоритмов поиска аномалий для текстовых коллекций, методики формирования стратегии выявления аномально новых документов, эксперимента по поиску оптимальных стратегий выявления аномально новых документов для двух наборов данных, описания результатов и возможностей их инженерного применения.

2. Обзор. Исследования по алгоритмам обнаружения аномалий активно ведутся с 1980-х годов [9]. Методы решения задачи поиска аномалий в текстовых коллекциях могут быть классифицированы различным образом. В работе [10] подходы к выявлению аномалий делятся на две категории: статистические и нейросетевые. В работе [11] различные методы анализируются в том числе с точки зрения применимости к той или иной прикладной задаче, определяемой типом исследуемых данных (изображения, тексты, медицинские сведения, банковские транзакции, биржевые сделки, статистика телефонных звонков и т.д.). Для анализа текстовых коллекций предлагается использовать такие методы, как моделирование на основе смеси распределений, статистическое профилирование с использованием гистограмм, метод опорных векторов, нейронные сети, кластеризацию.

В обзорной статье [12] рассматривается общая задача обнаружения новизны в коллекциях элементов без учета формата (типа данных) этих элементов, и выделяются следующие подходы к ее решению: вероятностное обнаружение аномальных элементов; методы на основе вычисления расстояний в пространстве коллекции; методы снижения размерности и последующей реконструкции; обнаружение аномалий с учетом знаний о предметной области. В [13] анализируются различные аспекты обнаружения новизны в потоке данных: автономная и онлайновая фазы алгоритма, количество классов, рассматриваемых на каждой

фазе, ансамблевые методы сравниваются с решением на основе единственного классификатора, рассматриваются подходы к обучению (с учителем и без него), а также вопросы обновления модели принятия решений и другие.

С ростом интереса к нейросетевым подходам классификация методов обнаружений аномальных элементов претерпела изменения. Так, авторы обзорной работы этого года [14] выделяют две обширные категории методов поиска аномалий в зависимости от коэффициентов нейросетевой модели (feature map): глубокие и неглубокие (deep и shallow), относя к последним и те методы, которые формально не являются нейросетевыми.

Современные методы обнаружения аномалий в текстовых данных сталкиваются с несколькими проблемами, включая сильную разреженность данных, зависимость от метрики расстояния пространства признаков, возникновение большого количества кластеров, а также неизвестные признаки аномальных элементов. Актуальность той или иной проблемы зависит от конкретной прикладной задачи, определяемой бизнес-процессом и особенностями текстового корпуса.

Полный алгоритм обнаружения аномалий в реальной прикладной информационной системе складывается из последовательности шагов, на каждом из которых решается определенная подзадача и допускается несколько вариантов методов ее решения. На рисунке 1 показана совокупность подзадач и методов, которые были рассмотрены в рамках настоящего исследования.

2.1. Тип задачи: поиск выбросов или обнаружение новизны. Большинство методов поиска аномалий вычисляют функцию оценки выброса (outlier score), характеризующую степень аномальности каждого объекта пространства, а также определяют пороговые значения этой оценки для обнаружения инородных элементов коллекции. Наиболее широко используемые методы пороговой обработки основаны на таких статистических данных, как стандартное отклонение от среднего, медианное абсолютное отклонение и межквартильный диапазон. К сожалению, при проверке новых элементов на аномальность по отношению к исходной коллекции такая статистика может оказаться необъективной, искажаясь за счет имеющихся выбросов. Минимизировать их влияние на вычисление оценки аномальности можно за счет предъявления дополнительных требований к коллекции и уточнения условий задачи: к однородной коллекции добавляется ровно один документ, который анализируется на предмет аномальности; за точку отсчета при оценке принимается центр коллекции. Такая постановка задачи рассматривается в настоящем исследовании.

2.2. Тип обучения: с привлечением учителя и без него. Для обнаружения аномалий в текстовой коллекции могут применяться разные подходы к обучению. Для решения этой задачи могут использоваться методы с учителем (supervised) либо с частичным привлечением учителя (semi-supervised) [1, 15]. Выявление аномалий в режиме без учителя также применяется [16], будучи востребованным в ситуациях, когда невозможно сообщить алгоритму предполагаемые характеристики аномальных документов или предоставить их образцы. Кроме того, такие методы не требуют трудоемкой разметки большого количества данных на основании экспертных знаний. Однако для задачи поиска несоответствующих элементов в коллекции объемных текстовых документов, которая рассматривается в рамках данного исследования, наиболее подходят методы обучения с частичным привлечением учителя, поскольку в реальных прикладных системах известны общие характеристики «корректных» документов, и у пользователей есть представление о том, какие документы точно не должны попасть в коллекцию.

2.3. Принцип формирования набора признаков. При поиске аномалий документы рассматриваются как объекты с некоторым набором признаков. Векторы признаков могут формироваться автоматически, при построении модели коллекции на основе статистических данных словаря, либо в результате процесса конструирования признаков (feature engineering) [17], когда для каждого из документов определяется набор характеристик, которые с точки зрения эксперта являются значимыми для разделения ядра коллекции и ее аномалий. Например, такими признаками могут быть именованные сущности, которые встречаются в тексте, или части речи каждого из слов [18]. При этом числовые значения признаков также могут быть вычислены в автоматическом режиме, в том числе на основе статистики словаря или характеристик графового представления текстов [19].

Плюс первого метода состоит в его универсальности и независимости от предметной области, что на практике означает широкую сферу применения алгоритма и отсутствие необходимости существенной перенастройки прикладной информационной системы для каждого заказчика. К преимуществам второго способа формирования признаков можно отнести умеренную вычислительную сложность и возможность максимальной адаптации под конкретную прикладную задачу.

Следует отметить, что автоматическое конструирование признаков без учета специфики обрабатываемых данных применимо не всегда. Например, метод COPOD (Copula-Based Outlier Detection) [20], демонстрирующий результативность на разнообразных наборах данных, опирается на набор признаков, сформированный на основании копулы -

функции от многомерного распределения, позволяющей отделить частные распределения от структуры зависимостей данного многомерного распределения. Существенным преимуществом этого метода является отсутствие параметров (в отличие от других методов, где выбор параметров способен существенно повлиять на результат) и высокая производительность, в том числе на пространствах большой размерности. Однако заявленная в статье эффективность метода не была экспериментально показана его авторами на текстовых корпусах и не подтвердилась в рамках настоящего исследования, поэтому метод был исключен из рассмотрения.

2.4. Метод построения векторной модели. Базовым вариантом автоматического конструирования набора признаков для коллекции текстовых документов является построение их векторного представления на основе статистических характеристик словаря коллекции. В основе такого представления лежит модель Bag Of Words («мешок слов», BoW), в которой порядок слов в исходных текстах не учитывается.

При отсутствии или неявной выраженности содержательной специфики коллекции документов пространство признаков может быть построено с помощью методов дистрибутивной семантики на основе уже известных данных о распределении слов в универсальных языковых корпусах (word2vec [21]).

Однако, если учитывать особенности лексики документов в рамках определенной прикладной системы, точность решения может оказаться существенно выше. Для реализации этого преимущества строится матрица «терм-документ», содержащая веса, вычисляемые как значения функции TF-IDF.

Еще одним способом построения векторной модели текстовой коллекции является тематическое моделирование. В этом случае для представления коллекции строятся две модели: «терм-тема» и «тема-документ», и вектор каждого документа содержит веса тем. Плюсом такого представления текстовой коллекции для обнаружения аномалий является гораздо меньшая, по сравнению с предыдущими двумя способами, размерность пространства признаков, а также более высокая объяснимость решения, что важно при разработке прикладных информационных систем. Однако процесс построения тематической модели требует больше вычислительных ресурсов и более чувствителен к выбору параметров алгоритма.

2.5. Метрика сравнения распределений. Необходимым этапом при решении задачи поиска аномалий в коллекции является выбор метрики аномальности нового документа (скоринговой функции, скоринга аномальности).

В решаемой задаче вычисление скоринга аномальности нового документа может быть сделано на основе сравнения расстояния от каждого документа исходной коллекции до двух объектов: до ее центрального элемента и до нового документа. Полученные два набора расстояний между документами сравниваются между собой с помощью метрики различия между распределениями.

В качестве такой метрики могут быть рассмотрены, например, расстояние Хеллингера, дивергенция Кульбака-Лейблера, перекрестная энтропия и другие функции. На этом этапе решения задачи необходимо подобрать наиболее выразительную метрику сравнения распределений для данной коллекции текстовых документов, которая и станет результирующей скоринговой функцией для рассматриваемых входных данных: коллекции документов и образца инородного документа.

2.6. Параметры токенизации текста. При построении векторной модели коллекции на основе статистики словаря возможны различные варианты токенизации текста: в качестве термов могут рассматриваться униграммы, *n*-граммы, субсловарные единицы (части слов) разной длины либо их комбинация. Выбор принципа токенизации влияет на вычислительную сложность алгоритма и одновременно на его точность. При этом для некоторых прикладных задач допустимо не приводить слова к начальной форме. Удаление из словаря наиболее часто и наиболее редко встречающихся слов также опционально, как и любая другая его фильтрация.

2.7. Метрика расстояния в векторном пространстве. В построенном векторном пространстве документов коллекции необходимо задать метрику расстояния между элементами. Близость между документами в таком представлении может быть измерена различными способами, наиболее объяснимыми среди которых являются Манхэттенское расстояние, Евклидово расстояние и косинусная мера. При этом оптимальной метрикой пространства при работе с векторным представлением текстовых данных является косинусная мера, что, в частности, продемонстрировано в работе [22].

2.8. Способ дообучения модели. При добавлении в коллекцию нового документа, который должен быть оценен на предмет аномальности по отношению к данной коллекции, для него необходимо построить вектор в имеющемся семантическом пространстве. В случае, если используются методы дистрибутивной семантики, вектор создается по тем же принципам, что и вся модель. Если же используется метод построения матрицы коллекции по TF-IDF или строится тематическая модель, то существуют два варианта создания вектора для нового документа.

Первый способ - построение вектора нового документа в имеющемся пространстве без увеличения его размерности. В этом случае словарь коллекции не меняется, поскольку новый документ не обогащает его. Второй способ - пересчет всей модели с учетом добавления в коллекцию нового документа. В этом случае словарь расширяется за счет слов нового документа.

При решении реальных задач в рамках прикладных информационных систем выбор варианта построения вектора нового документа определяется имеющимися вычислительными ресурсами и необходимой скоростью, а также предполагаемым сценарием работы с новым документом. При последующем сохранении нового документа в системе полный пересчет модели является оправданным решением, а если аномальный документ удаляется сразу после анализа, то включать его слова в словарь модели нецелесообразно, и вектор должен рассчитываться по существующему словарю.

2.9. Методы снижения размерности векторов. При формировании векторных представлений текстовых коллекций на основе статистики их словаря строятся пространства большой размерности, при этом векторы документов получаются сильно разреженными. Это затрудняет работу многих алгоритмов определения аномалий, делая их неэффективными для анализа текстовых данных. Такие алгоритмы требуют предварительной процедуры понижения размерности матриц.

Уменьшение размерности с помощью линейных алгоритмов, таких как метод главных компонент (PCA) и разложение по сингулярным числам (TruncatedSVD), не приносит большой пользы из-за того, что эти алгоритмы линейные. Даже сохраняя в модели большую часть разнообразия данных, при сворачивании измерений они не способны к нелинейным преобразованиям, которые могут учитывать информацию о совстречаемости слов.

В то же время алгоритмы, основанные на нелинейных преобразованиях (такие как SNE и его модификации [23]) не сохраняют отношения плотности и расстояний. Однако в случае разреженных данных сохранение локальных областей с высокой плотностью приобретает особую важность. Это способны делать некоторые нелинейные алгоритмы, например, UMAP [24]. Алгоритмы denSNE и densMAP [25] могут сохранять информацию о плотности при правильном выборе параметров: они вычисляют оценки локальной плотности и используют эти оценки в качестве регуляризатора при оптимизации низкоразмерного представления.

2.10. Алгоритмы выявления аномалий. После построения векторной модели коллекции и установления необходимых метрик следующим этапом в решении задачи является выбор метода поиска аномальных элементов.

В рамках данного исследования изучается применимость алгоритмов, относящихся, согласно классификации [14], к категории неглубоких. В качестве базового требования к алгоритмам, исходящего из реального опыта разработки прикладных систем, авторами была установлена достаточно хорошая объяснимость получаемых решений. По этой причине не рассматривались нейросетевые типы алгоритмов, относящиеся в указанном обзоре к категории deep. Кроме того, исключались из рассмотрения вероятностные алгоритмы, что обусловлено требованиями воспроизводимости результатов в реальных прикладных информационных системах.

В таблице 1 для различных групп методов машинного обучения показано сопоставление их объяснимости и точности, а также указана их вычислительная сложность.

Таблица 1. Общие характеристики методов обнаружения аномалий

Метод	**Точность**	**Объяснимость**	**Сложность вычислительная**
Логистическая регрессия	Низкая	Высокая	$O(n \cdot d)$, где n - количество элементов, d - размерность; используется для данных низкой размерности.
Дерево решений	Низкая	Высокая	$O(n \cdot \log(n) \cdot d)$ - вычислительная сложность обучения модели; $O(n)$ - сложность во время выполнения; используется для данных низкой размерности.
Метод ближайших соседей (локальный фактор выброса)	Средняя	Средняя	$O(k \cdot n \cdot d)$, где n - количество элементов, d - размерность, k - количество соседей.
Кластеризация	Средняя	Средняя	Зависит от алгоритма кластеризации: $O(n \cdot k \cdot l)$, где k – число кластеров, l – число итераций для k-means.

Продолжение таблицы 1.

Метод	**Точность**	**Объяснимость**	**Сложность вычислительная**
Метод опорных векторов	Средняя	Средняя	$O(n^2)$ или $O(n^3)$ - вычислительная сложность обучения модели в зависимости от используемого ядра; $O(k \cdot d)$ - сложность во время выполнения, где k - количество опорных векторов, d - размерность данных.
Ансамблевые методы (случайный лес)	Высокая	Низкая	$O(n \cdot \log(n) \cdot d \cdot k)$ - вычислительная сложность обучения модели, где k - количество деревьев; $O(n \cdot k)$ - сложность во время выполнения.
Нейронные сети	Высокая	Низкая	$O(n^4)$ – прямое распространение; $O(n^5)$ – обратное распространение ошибки.

Исходя из требований реальной задачи – точности решений при достаточно высоком уровне их объяснимости и воспроизводимости (что на практике означает надежность рекомендаций информационной системы) – в рамках данной работы для построения оптимальной стратегии выявления аномалий в коллекции документов рассматриваются следующие методы: локальный уровень выброса, мягкая кластеризация, метод опорных векторов, а также один из ансамблевых методов – изолирующий лес.

Метод ближайших соседей характеризуется наиболее высокой объяснимостью, поскольку интуитивно понятно, что в векторном пространстве у аномалии мало близких соседей, а у типичной точки (относящейся к т. н. ядру коллекции) их много. Поэтому величина «расстояние до *k*-го соседа» может служить мерой аномальности документа. Метод хорошо работает на однородных коллекциях, позволяющих задать единый порог аномальности для всех элементов коллекции, либо на коллекциях, состоящих из нескольких кластеров, плотность которых примерно равна.

В то же время для менее однородных коллекций значимым может оказаться обнаружение локальных аномалий, то есть выбросов, в том

числе для задачи приведения коллекции к однородному состоянию. Метод локального фактора выброса (Local Outlier Factor, LOF) позволяет присвоить каждому объекту степень локальной аномальности (локальный фактор выброса). Она показывает, насколько объект изолирован от других в окружающей его окрестности, то есть от совокупности его ближайших соседей. Преимущество LOF перед другими алгоритмами обнаружения аномалий проявляется в способности обрабатывать наборы данных с кластерами разной плотности.

Результаты работы алгоритмов зависят не только от выбранных вариаций и параметров, но и от свойств набора данных, на которых этот алгоритм определения аномалий применяется. В работе [26] экспериментально показано, что алгоритм ближайших соседей по-разному работает на различных дата-сетах, хотя все они репрезентативны и созданы специально для исследования задачи выявления аномалий.

Методы кластеризации определяют аномалии как элементы, которые не могут быть включены ни в один из обнаруженных кластеров. Кластеризация может проводиться по различным критериям: по тематической принадлежности элементов, по плотности их распределения, на основании дополнительных признаков.

Кластеризация по плотности распределения [27] для текстовых документов неэффективна из-за большой размерности пространства в совокупности с сильной разреженностью векторных представлений. В матрицах модели BoW доля ненулевых значений составляет около 1% и даже менее, что затрудняет формирование кластеров. Поэтому такие алгоритмы кластеризации, как HDBSCAN [28], OPTICS [29], неприменимы для полноразмерных векторных представлений текста.

Алгоритм кластеризации CHAMELEON [30], основанный на сегментации графа методами семейства METIS [31], также подвержен «проклятию размерности». Алгоритм работает со связными графами, тогда как по разреженному матричному представлению формируются многочисленные связные компоненты малого размера, и анализ такого графа становится неэффективным.

В работе [32] для определения аномалий используется алгоритм кластеризации с помощью модификации неотрицательного матричного разложения (Non-negative Matrix Factorization, NMF), и аномальными считаются документы, которые не соответствуют построенной тематической модели.

Кроме того, кластеры сами могут быть основой для формирования набора признаков: например, в работе [33] описана вероятностная кластеризация, и характеристики кластеров передаются классификатору в качестве дополнительных признаков элементов коллекции.

Метод опорных векторов (Support Vector Machine, SVM) может рассматриваться в задаче обнаружения аномалий как метод кластеризации с единственным кластером. Этот подход реализован в алгоритме One-Class SVM [34]. Будучи чувствительным к выбросам, он лучше подходит для обнаружения новизны, когда обучающая выборка достаточно однородна. Тем не менее, обнаружение выбросов в пространстве большой размерности, а также без каких-либо предположений о распределении входящих данных является очень сложной задачей, и One-Class SVM может оказаться полезным, если выбрать подходящие значения его гиперпараметров.

Ансамблевые методы поиска аномалий как правило предполагают обучение с привлечением учителя, хотя ансамблям без учителя посвящены некоторые исследования [35]. Чтобы автоматизировать оценку эффективности различных ансамблевых методов, в работе [36] предлагается использовать скоринговые функции.

К ансамблевым методам относится метод изолирующего леса (Isolation Forest, iForest [37]), реализованный на основе ансамбля Extra Tree Regressor. В авторском варианте максимальная глубина каждого дерева устанавливается равной количеству сэмплов, использованных для построения дерева, что увеличивает скорость работы метода. Кроме того, iForest допускает распараллеливание при вычислении, что также ускоряет их. К плюсам этого метода также относится его способность эффективно работать с мультимодальными наборами данных; то же характеризует и метод LOF.

Обзор большого числа научных работ с описанием методов и модификаций, демонстрирующих хорошую точность и высокую эффективность, позволяет сделать вывод, что результаты достоверны преимущественно лишь на описываемых экспериментальных данных либо на аналогичных им. В частности, не все методы тестируются на текстовых корпусах, в том числе на русскоязычных - большинство экспериментов проводится с данными на английском языке. Поэтому, несмотря на результативность представленных во многих работах алгоритмов, они не могут быть непосредственно применены для анализа русскоязычных документов [38], для этого требуется их адаптация с учетом специфики русского языка. Кроме того, очевидно, что такие характеристики корпуса, как длина текстов, широта тематики, структурированность документов и однородность коллекции имеют большое значение при выборе алгоритма поиска аномалий.

3. Методы. В рамках данной работы изучалась возможность выбора оптимального алгоритма для обнаружения аномальности нового

документа по отношению к однородной коллекции русскоязычных текстовых документов средней и большой длины, обладающих юридической значимостью.

Для решения задачи была выбрана следующая стратегия: среди различных алгоритмов обнаружения аномалий подобрать оптимальный для конкретной коллекции документов, с учетом циклического перебора гиперпараметров на каждом шаге алгоритма и наличия образца аномального документа.

Рассматривались методы поиска аномалий из числа достаточно объяснимых и способных обеспечивать приемлемую точность. Первоначальный выбор методов был обусловлен накопленными у авторов сведениями об их объяснимости, характеристиками вычислительной сложности, а также известной точностью результатов. Исходя из этих требований нейросетевые методы не рассматривались. Таким образом, для исследования были выбраны следующие методы: локальный фактор выброса (LOF), изоляционный лес (iForest), метод опорных векторов для одного класса (One-Class SVM).

Для целей исследования была проведена серия экспериментов. В качестве экспериментальных данных были выбраны две однородные коллекции нормативных документов, представляющих собой структурированные тексты большой и средней длины, схожие по тематике. Векторное представление коллекции документов формировалось двумя способами: на основе функции TF-IDF и методом тематического моделирования.

Для выбора метрики был использован следующий подход в рамках обучения с привлечением учителя: к исходной коллекции добавляется априори инородный документ и среди различных метрик расстояния между распределениями выбирается такая, которая показывает различие между ядром коллекции и инородным элементом максимально «контрастно». На основе этой метрики, оптимальной для данной коллекции, может быть построена метрика аномальности.

В ходе предварительного анализа в качестве метрики расстояния между распределениями было выбрано расстояние Хеллингера, которое рассчитывается по формуле (1).

$$H(P,Q) = \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{i=1}^{k} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2}\ , \qquad (1)$$

где $P = (p_1, ..., p_k)$ – нормированный вектор косинусных расстояний от всех документов коллекции до ее центра, $Q = (q_1, ..., q_k)$ – нормированный вектор косинусных расстояний от всех документов коллекции до инородного документа.

Для обеих коллекций был выбран набор образцов аномальности – инородных документов разного типа. Функция оценки аномальности устанавливалась в зависимости от алгоритма, однако общий принцип ее вычисления базировался на сравнении расстояний от нового элемента до двух точек: до центра коллекции и до одного из инородных документов. Центром коллекции считается документ, сумма расстояний от которого до всех остальных документов минимальна (по косинусной метрике).

Для каждого из рассматриваемых алгоритмов был проведен процесс подбора оптимальных параметров. В качестве критериев оптимизации выступали следующие величины: количество документов коллекции, оценка аномальности которых выше, чем у инородного документа; расстояние от центра коллекции до инородного документа. Таким образом, целью оптимизации было получение таких параметров, при которых инородные документы оказались бы элементами с наиболее высокой оценкой аномальности (с учетом условия однородности коллекции), и при этом картина распределения оценок аномальности была бы наиболее выразительная.

4. Эксперимент. Цель эксперимента состояла в проверке стратегии обнаружения аномалий, приведенной в разделе «Методы», на экспериментальных данных. Для достижения поставленной цели были выделены этапы:

1. Подготовка набора данных для проведения эксперимента.
2. Выбор метода построения векторной модели.
3. Выбор параметров построения словаря векторной модели, которая обеспечивает наибольшее расстояние от центра коллекции до инородного документа.
4. Подбор методов обнаружения новизны/выбросов для векторной модели документов, полученной с помощью преобразования TF-IDF.
5. Подбор методов обнаружения новизны/выбросов для векторной модели документов, полученной с помощью тематического моделирования.

Для проведения эксперимента было выбрано 2 коллекции юридически значимых документов ГОСТ:

- относящихся к тематике информационных технологий (далее - ГОСТ ИТ) – 1198 документов[1];
- относящихся к тематике железных дорог (далее - ГОСТ ЖД) – 458 документов[2].

В качестве образцов аномальности для вышеуказанных коллекций было выбрано 5 инородных документов (далее - ИД), которые являются аномальными для данных коллекций согласно экспертной оценке, список документов и фрагменты текстов (для примера) приведены в таблицах 2 и 3.

Таблица 2. Образцы аномальности - инородные документы[3]

Обозначение в эксперименте	**Категория**	**Название ИД**	**Длина документа**
ИД0	Научная статья	Цифровой двойник сортировочной горки	1505 слов
ИД1	Финансовая отчетность	Заключение по результатам обзорной проверки промежуточной финансовой информации ОАО «РЖД» и его дочерних компаний	14842 слова
ИД2	Статья в СМИ	Сколько строить и в чем возить: транспортное машиностроение на пространстве	1712 слов
ИД3	Шаблон договора	Договор с ОАО «РЖД» на оказание услуг, связанных с перевозкой грузов	4327 слов
ИД4	Художественная литература	Отрывок из романа Л. Н. Толстого «Война и Мир» (т.1, ч.1, гл. I-III)	4815 слов

[1] Документы из раздела 35. Информационные технологии. Машины конторские в Общероссийском классификаторе стандартов (ОКС)

[2] Документы из раздела 45. Железнодорожная техника в ОКС

[3] Все перечисленные документы размещены в открытом доступе в Интернете.

Таблица 3. Фрагменты инородных документов

Обозначение	**Название документа**
Фрагмент текста документа	
ИД0	Цифровой двойник сортировочной горки
В целях моделирования наиболее оптимальных режимов работы объекта автоматизации (в зависимости от конкретной ситуации и выбранного критерия оптимизации и анализа возможности оптимизации численности персонала, обслуживающего и управляющего объектом) ведется разработка цифровых двойников как элементов станционной инфраструктуры, так и сортировочной горки в целом – как наиболее сложной в плане автоматизации процессов и обеспечения безопасности части сортировочной станции. Создание цифровых двойников элементов железнодорожной инфраструктуры и подвижного состава является логичным развитием концепции цифровой станции и технологии промышленного интернета вещей. Цифровой двойник (digital twin) – это перенесенный в цифровую среду двойник физического устройства, процесса или системы. Цифровой двойник – это математическая модель высокого уровня адекватности, которая позволяет с большой точностью описывать поведение объекта во всех ситуациях, на всех этапах жизненного цикла, включая аварийные. Применение цифровых двойников позволяет быстро смоделировать развитие событий в зависимости от тех или иных факторов, определить потенциальные риски, найти наиболее эффективные режимы работы, выстроить шаги по обеспечению безопасности.	
ИД1	Заключение по результатам обзорной проверки промежуточной финансовой информации ОАО «РЖД» и его дочерних компаний
В состав сумм в таблице выше по состоянию на 30 июня 2020 г. включены суммы по договорам, заключенным со связанными сторонами – совместными предприятиями Компании и компаниями под контролем Российской Федерации, в размере 36 503 миллиона рублей и 68 328 миллионов рублей, соответственно. По состоянию на 30 июня 2020 г. обязательства по данным договорам со связанными сторонами составили 7 496 миллионов рублей и 36 424 миллиона рублей, соответственно. Договоры на приобретение локомотивов, заключенные в 2018–2019 годах, также предусматривают обязательства по обеспечению сервисного обслуживания в период жизненного цикла на срок до 28 лет, не включенные в суммы в таблице выше. Кроме того, в 2017–2020 годах Группа заключила договоры на выполнение работ по капитальному ремонту и модернизации подвижного состава, капитальному ремонту объектов электрификации и электроснабжения на общую сумму 170 950 миллионов рублей (в том числе 123 397 миллионов рублей по договорам, заключенным со связанными сторонами – совместными предприятиями Компании по состоянию на 30 июня 2020 г.). Стоимость работ, планируемых к выполнению по данным договорам после 30 июня 2020 г., составляет 105 873 миллиона рублей (в том числе 66 315 миллионов рублей по договорам, заключенным со связанными сторонами по состоянию на 30 июня 2020 г.).	

Продолжение таблицы 3.

Обозначение	**Название документа**
Фрагмент текста документа	
ИД2	Сколько строить и в чем возить: транспортное машиностроение на пространстве
Разбалансировка рынка Большинство железнодорожных компаний стран пространства являются крупнейшими системообразующими элементами экономики и ключевыми звеньями транспортных систем своих государств. В России 2010–2019 гг. были достаточно плодотворными с точки зрения появления современной продукции для транспортного машиностроения. Правда, в последнее время из-за перенасыщения рынка грузовых вагонов динамика производства в отечественном железнодорожном машиностроении значительно ухудшилась. Такая тенденция наблюдается уже не первый год, а сейчас на нее наложились еще и последствия пандемии коронавируса, из-за которых произошло резкое снижение грузовых перевозок и сокращение инвестиционных планов транспортных компаний. По данным РЖД, всего в этом году планируется модернизировать 449 вагонов. Кроме того, 2 октября на Павелецком вокзале Москвы состоялась презентация нового концепта плацкартного пассажирского вагона для поездов дальнего следования. На фоне пандемии вагоностроители столкнулись с серьезным падением спроса и прогнозами на его замораживание в среднесрочной перспективе. Заместитель генерального директора ИПЕМ Владимир Савчук поясняет, что снижение спроса на новые грузовые вагоны в России прогнозировалось заранее и максимальные риски среди заводов были у НПК «Уралвагонзавод» (УВЗ, входит в «Ростех») из-за концентрации линейки выпускаемого предприятием подвижного состава в наиболее рискованных сегментах рынка – полувагонах и нефтебензиновых цистернах.	
ИД3	Договор с ОАО «РЖД» на оказание услуг, связанных с перевозкой грузов
5. Ответственность Сторон 5.1. Стороны несут ответственность за неисполнение или ненадлежащее исполнение своих обязательств по настоящему Договору в соответствии с действующим законодательством Российской Федерации. 5.2. За просрочку платежей по настоящему Договору ОАО «РЖД» вправе предъявить Клиенту требования об уплате пени в размере 0,1 % от суммы задолженности за каждый день просрочки. 5.3. В случае отсутствия в ОАО «РЖД» уведомления об изменении местонахождения, почтового адреса, других реквизитов Клиента, учредительных, уставных документов, а также непредставления Клиентом заверенных надлежащим образом копий подтверждающих документов в сроки, установленные подпунктами 9.1 и 9.2 настоящего Договора, ОАО «РЖД» вправе приостановить выполнение своих обязательств по настоящему Договору до предоставления Клиентом указанных сведений и документов. 5.4. Исполнитель по настоящему Договору не является Грузоотправителем и не несет ответственность, предусмотренную законодательством РФ как Грузоотправитель.	

Продолжение таблицы 3.

Обозначение	**Название документа**
Фрагмент текста документа	
ИД4	Отрывок из романа Л. Н. Толстого «Война и Мир»
Гостиная Анны Павловны начала понемногу наполняться. Приехала высшая знать Петербурга, люди самые разнородные по возрастам и характерам, но одинаковые по обществу, в каком все жили; приехала дочь князя Василия, красавица Элен, заехавшая за отцом, чтобы с ним вместе ехать на праздник посланника. Она была в шифре и бальном платье. Приехала и известная, как la femme la plus séduisante de Pétersbourg,[34] молодая, маленькая княгиня Болконская, прошлую зиму вышедшая замуж и теперь не выезжавшая в большой свет по причине своей беременности, но ездившая еще на небольшие вечера. Приехал князь Ипполит, сын князя Василия, с Мортемаром, которого он представил; приехал и аббат Морио и многие другие. — Вы не видали еще? или: — вы не знакомы с ma tante? [35] — говорила Анна Павловна приезжавшим гостям и весьма серьезно подводила их к маленькой старушке в высоких бантах, выплывшей из другой комнаты, как скоро стали приезжать гости, называла их по имени, медленно переводя глаза с гостя на ma tante,[36] и потом отходила.	

Эксперимент по определению оптимальных характеристик словаря векторной модели проводился для обеих коллекций ГОСТов, в качестве ИД был выбран наиболее отличающийся от них документ ИД4 – сопоставимый по объему отрывок из романа «Война и Мир».

Для подбора оптимальной метрики аномальности были выбраны следующие варианты построения словаря векторной модели документов:

- субсловарные термы (*n*-граммы из 2–4 символов);
- субсловарные термы (*n*-граммы из 2–4 символов) с редуцированием словаря (РС): удалены русские стоп-слова, часто встречающиеся и редкие термы;
- униграммы и биграммы;
- униграммы и биграммы с РС.

Векторизация документов осуществлялась в двух вариантах: с помощью алгоритма расчета TF-IDF и на основе количества вхождений слова в документ. Векторное преобразование TF-IDF осуществлялось с нормализацией L1 (нормализация вектора на сумму абсолютных значений всех его компонентов). Для токенизации применялись два метода: TweetTokenizer (далее - TT) и RegexpTokenizer (далее - RT). Результаты расчета расстояния Хеллингера между центром коллекции и документом ИД4 представлены в таблице 4.

Таблица 4. Значения расстояния Хеллингера между центром коллекции и документом ИД4 (отрывок из «Войны и мира») при различных вариантах построения векторной модели

Словарь векторной модели	Коллекция: ГОСТ ИТ				Коллекция: ГОСТ ЖД			
	TF-IDF		Счетчик слов		TF-IDF		Счетчик слов	
	TT	RT	TT	RT	TT	RT	TT	RT
Субсловарный	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Субсловарный с РС	0.04	0.04	0.03	0.05	0.06	0.06	0.07	0.07
Униграммы и биграммы	0.03	0.03	0.05	0.05	0.04	0.04	0.09	0.1
Униграммы и биграммы с РС	0.03	0.03	0.05	0.05	0.04	0.04	0.05	0.05

Из таблицы видно, что наилучший результат (максимальное расстояние между распределениями) для обеих коллекций показала метрика аномальности на основе расстояния Хеллингера для субсловарных 2–4-грамм без редуцирования словаря по частотам. Дальнейший подбор методов выявления аномалий в текстовых коллекциях проводился при этих параметрах словаря.

В рамках подбора методов и параметров для обнаружения аномальных данных был проведен ряд экспериментов с применением следующих алгоритмов:

- локальный фактор выброса (LOF): алгоритм – поиск методом перебора, количество соседей – 2. Метод проверялся для разных значений метрики расстояния: cosine и L2;
- изоляционный лес (iForest): количество деревьев (эстиматоров) – 500;
- метод опорных векторов для одного класса (One-Class SVM): в качестве функции преобразования признакового пространства-ядра (kernel) рассматривались линейная функция (linear)/ сигмоида (sigmoid)/ радиальная базисная функция (RBF).

Использовалась реализация алгоритмов в Python-библиотеке Scikit-learn.

Указанные алгоритмы применялись для векторов документов, сформированных с помощью векторного преобразования TF-IDF для субсловарных 4-грамм без редуцирования словаря по частотам.

Для каждого из алгоритмов было исследовано влияние размерности векторов документов на результат работы. Проверялась работа алгоритма без снижения размерности, а также с применением следующих методов ее снижения:

- нелинейный метод снижения размерности UMAP: размерность конечного пространства – 300, метрика расстояния – косинусное расстояние;
- линейный метод на основе сингулярного разложения (Truncated SVD: размерность конечного пространства – 300). Метод тестировался при двух вариантах алгоритма разложения: итерационным методом Арнольди (SVD arpack) и ускоренным алгоритмом Халко (SVD random)).

Обнаружение аномалий проводилось каждым из алгоритмов поочередно в двух режимах, соответствующих виду задачи: поиск выбросов (обучение модели на коллекциях документов с добавленными инородными документами) и обнаружение новизны (обучение модели на коллекциях документов без добавленных инородных документов).

После получения функции принятия решения (decision-функции) по каждому алгоритму была произведена оценка качества ее работы: были определены документы коллекции, которые decision-функция отнесла к аномалиям с индексом выше, чем ИД2 (статья в СМИ). Результаты анализа представлены в таблице 5.

Из таблицы видно, что наилучший результат для обеих коллекций получен на полноразмерных векторах документов без использования методов снижения размерности для задачи обнаружения новизны. При таком варианте найдено наименьшее количество документов коллекции, отнесенных к аномальным с индексом выше, чем инородный документ. Также было обнаружено, что в данном варианте нет документов с индексом аномальности выше, чем у художественного текста (см. рис. 2). Наихудший результат получен при использовании нелинейного метода снижения размерности UMAP.

Наиболее стабильный результат для обеих коллекций вне зависимости от выбора метода снижения размерности показал алгоритм, основанный на методе опорных векторов (One-Class SVM). Работа метода была проверена на моделях обеих коллекций, построенных на векторах документов, сформированных с помощью преобразования TF-IDF для словаря из униграмм и биграмм с редуцированием по частотам (удалены часто встречающиеся и редкие термы) – такая векторная модель наиболее часто используется в ПИИС.

Эксперимент показал, что метод One-Class SVM с функцией преобразования признакового пространства RBF лучше определяет аномалии при нормализации векторной модели по L2, при которой вектор нормализуется на сумму квадратов всех его значений (см. табл. 6).

Также было обнаружено, что наилучшие результаты получены при детектировании инородных документов как новых для коллекций.

Таблица 5. Число документов, отнесенных к аномальным с индексом выше, чем у документа ИД2 (статья в СМИ), различными алгоритмами на основе полноразмерной векторной модели

Снижение размерности	Вид задачи	iForest	LOF		One-Class SVM		
			cosine	L2	sigmoid	linear	RBF
Коллекция: ГОСТ ИТ							
Без снижения	Выбросы	13	460	496	2	2	291
	Новизна	31	3	5	2	2	292
UMAP	Выбросы	1029	792	863	1102	593	818
	Новизна	1036	123	117	1102	593	818
SVD arpack	Выбросы	170	902	897	2	2	380
	Новизна	146	9	11	2	2	380
SVD random	Выбросы	156	881	864	2	2	384
	Новизна	182	7	10	2	2	384
Коллекция: ГОСТ ЖД							
Без снижения	Выбросы	12	20	36	5	5	202
	Новизна	7	2	1	5	5	200
UMAP	Выбросы	443	312	305	458	156	417
	Новизна	442	255	246	458	156	416
SVD arpack	Выбросы	112	20	41	5	5	194
	Новизна	151	2	3	5	5	192
SVD random	Выбросы	129	18	41	5	5	194
	Новизна	136	2	2	5	5	193

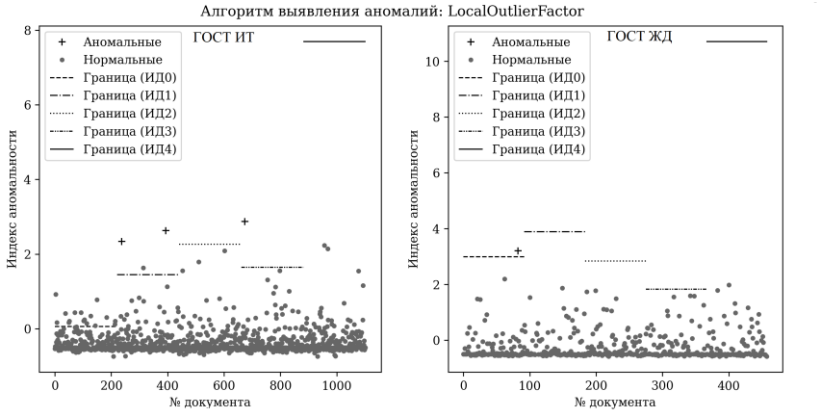


Рис. 2. Индекс аномальности документов в коллекциях ГОСТ ИТ и ГОСТ ЖД. Алгоритм Local Outlier Factor, векторная модель TF-IDF без понижения размерности

Таблица 6. Число документов, отнесенных к аномальным с индексом выше, чем у документа ИД2 (статья в СМИ), алгоритмом One-Class SVM на основе полноразмерной векторной модели

Снижение размерности	**Вид задачи**	**One-Class SVM (L1)**			**One-Class SVM (L2)**		
		sigmoid	**linear**	**RBF**	**sigmoid**	**linear**	**RBF**
Коллекция: ГОСТ ИТ							
Без снижения	Выбросы	4	110	51	8	8	6
	Новизна	0	0	51	3	2	0
SVD arpack	Выбросы	3	159	53	10	9	272
	Новизна	0	0	52	3	3	247
SVD random	Выбросы	0	136	53	10	9	264
	Новизна	2	0	53	5	4	242
Коллекция: ГОСТ ЖД							
Без снижения	Выбросы	125	386	5	3	3	1
	Новизна	0	0	4	0	0	0
SVD arpack	Выбросы	117	366	5	3	3	2
	Новизна	0	0	4	0	0	0
SVD random	Выбросы	118	364	4	3	3	2
	Новизна	0	0	5	0	0	0

Поскольку в ПИИС в зависимости от задачи могут использоваться разные способы векторного преобразования документов, то дальнейшие эксперименты проводились для векторной модели документов, полученной с помощью тематического моделирования.

Для обеих коллекций векторная модель документов строилась с методом тематического моделирования с аддитивной регуляризацией (Additive Regularization Topic Modelling, ARTM). При формировании словаря тематической модели была учтена информация о совстречаемости слов. Далее из 5 инородных документов была сформирована матрица «документ-тема» путем построения векторов на основе имеющихся тем.

Для получения decision-функции были проверены те же методы, что и для векторной модели текста TF-IDF: iForest, LOF, One-Class SVM. Для оценки качества работы функции также были определены документы коллекции, которые были отнесены к аномалиям с индексом выше, чем ИД2 (статья в СМИ). Результаты анализа представлены в таблице 7.

Таблица 7. Число документов, отнесенных к аномальным с индексом, превышающим индекс аномальности документа ИД2 (статья в СМИ), различными алгоритмами на основе тематической модели коллекции

Вид задачи	**iForest**		**LOF**		**One-Class SVM**	
	estimators: 400	**estimators: 500**	**cosine**	**L2**	**sigmoid**	**linear**
Коллекция: ГОСТ ИТ						
Выбросы	0	2	64	68	191	80
Новизна	0	0	6	97	191	77
Коллекция: ГОСТ ЖД						
Выбросы	0	0	339	401	305	27
Новизна	0	0	18	425	98	17

Из таблицы видно, что определение аномалий для тематической модели структурированных нормативно-технических документов на русском языке наилучшим образом произведено с помощью алгоритма изолирующего леса (Isolation Forest), при использовании этого метода все 5 ИД идентифицируются как аномальные: практически нет документов, превышающих ИД по индексу аномальности (см. рис. 3).

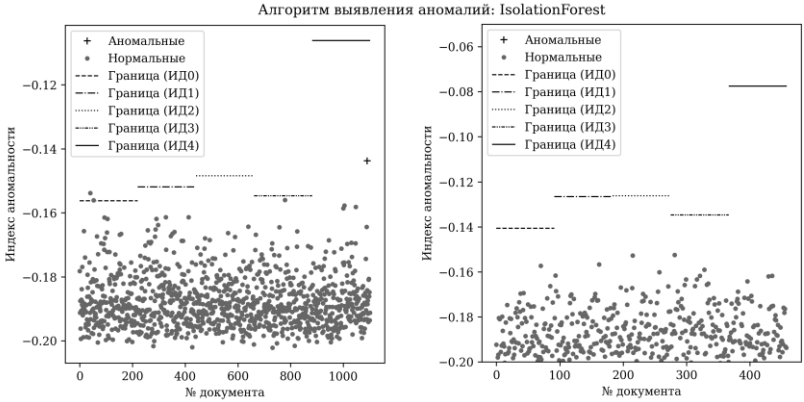


Рис. 3. Индекс аномальности для коллекций ГОСТ ИТ и ГОСТ ЖД. Алгоритм Isolation Forest, тематическая векторная модель (ARTM)

При использовании тематических векторов коллекций инородные документы изолируются алгоритмом гораздо раньше, чем нормальные объекты (выбросы попадают в листья на небольшой глубине дерева).

5. Результаты. Результаты эксперимента продемонстрировали эффективность предложенного подхода. При поиске аномалий в однородных текстовых коллекциях подбор параметров оптимизационным способом обеспечивает максимальную точность решения для конкретного набора данных с учетом предоставленного алгоритму образца инородного документа.

При этом выбор алгоритма зависит от целей и задач ПИИС: если для документов предполагается строить только векторную модель по TF-IDF, то выбор метода определяется большой размерностью пространства признаков. Как показал эксперимент, в этом случае снижение размерности нецелесообразно: наиболее эффективным оказались методы One-Class SVM (на основе метода опорных векторов) и LOF (локальный фактор выброса), работающие с полноразмерными векторами в режиме обнаружения новизны. Наиболее стабильный результат для обеих тестовых коллекций вне зависимости от выбора метода снижения размерности показал алгоритм One-Class SVM.

Эксперимент также показал, что наилучший результат можно получить при формировании словаря из субсловарных единиц в виде 2–4-грамм. Однако работа метода была также проверена при векторизации преобразованием TF-IDF по словарю из униграмм и биграмм с редуцированием по частотам - такая векторная модель наиболее часто используется в ПИИС. Экспериментально была установлена зависимость между выбором ядра и способом нормализации: линейное и сигмоидное ядра в One-Class SVM лучше работают при нормализации L1, тогда как преобразование пространства RBF эффективно при нормализации L2.

В случае, если в рамках ПИИС строится тематическая модель, аномалии эффективно определяются методом iForest (изолирующий лес), причем результат оказывается даже лучше, чем у оптимального алгоритма при векторизации методом TF-IDF.

Тематическое моделирование может быть рассмотрено как метод снижения размерности пространства: вместо векторов размерностью десятки тысяч компонентов (в соответствии со словарем коллекции) пространство формируется из векторов документов размерностью в несколько сотен тематических компонентов (что соответствует общепринятому порядку оптимального числа тем в тематической модели). Другие методы снижения размерности приводят к потере вместе с избыточной размерностью и некоторой части значимой информации - в частности, семантических признаков текстовых документов. Тематическое же моделирование позволяет сохранить эту информацию, поскольку уменьшение размерности модели происходит не за счет уплотнения векторов, а благодаря разложению исходной матрицы «документ-

терм» на две матрица меньшего размера - «терм-тема» и «тема-документ».

Тематическая модель также позволяет обнаружить аномальные элементы в процессе кластеризации. Аномалиями могут считаться документы, которые, несмотря на регуляризацию, «плохо» раскладываются по полученной модели: для них не обнаруживается подходящей темы с достаточным весом, вместо этого вектор такого документа содержит низкие значения весов по многим темам коллекции. Такие документы могут рассматриваться как выбросы. Однако при таком подходе возникает необходимость эмпирического выбора величины порога, которая в общем случае зависит от свойств коллекции. Пороговое значение необходимо выбирать таким образом, чтобы соблюсти баланс между числом кластеризуемых документов и числом документов, определяемых как аномальные. Кроме того, на практике этот подход невозможен для обнаружения новизны, поскольку полное переобучение тематической модели при поступлении каждого нового документа в ПИИС нецелесообразно.

Если же векторизовать новый документ по имеющемуся набору тем, то для обнаружения аномалий необходимо формализовать алгоритм принятия решения относительно качества разложения конкретного документа по имеющейся тематической модели. Это и делают алгоритмы обнаружения аномалий, проверенные экспериментально в данной работе. Примечательно, что стратегия поиска аномалий на основе тематической модели ARTM оказалось наиболее эффективной, при применении методов снижения размерности к модели TF-IDF таких результатов добиться не удавалось.

Таким образом, эксперимент подтвердил, что задача выявления единичных аномальных документов по отношению к заданной коллекции не имеет универсального решения. Однако возможно применение различных стратегий выявления аномалий в зависимости от потребности ПИИС.

Предложенный оптимизационный подход показал высокую эффективность при поиске смысловых аномалий на исследуемых данных. При этом необходимо отметить существующие ограничения текущего исследования. Для достижения высокой точности при поиске аномалий необходимо использовать знания о типе и структуре данных и об их предметной области. Это затрудняет оценку и сравнение с результатами других исследований: насколько нам известно, в настоящее время нет опубликованных исследований, посвященных поиску аномалий на уровне документов в коллекциях юридически значимых русскоязычных текстов. Предложенный подход оптимизации алгоритма обнаружения

аномалий с использованием заведомо инородного документа является новым, что также является причиной невозможности сопоставления полученного результата с предыдущими работами. Опубликованные исследования, посвященные обнаружению смысловой аномальности (новизны) на уровне документов, демонстрируют, что эффективность предлагаемых методов сильно зависит от обрабатываемого корпуса документов. В частности, в работе [2] приводятся оценки точности предложенного нейросетевого метода определения семантической новизны на уровне 0,75, 0,76 и 0,86 в зависимости от набора данных. Оптимизационный подход, представленный в данной статье, позволяет добиться абсолютной точности (близкой к 100%) при достаточном уровне однородности коллекции. Такая разница в эффективности демонстрирует зависимость оценок методов от качества тестовых данных. Это позволяет утверждать, что в прикладных задачах этап подготовки данных и выбор самого алгоритма поиска аномалий могут оказывать сопоставимо сильное влияние на результат анализа. В то же время при определении аномальности на семантическом уровне важно разграничивать случаи, когда ошибка вызвана работой системы (алгоритма), и ситуации ошибки из-за недостаточной формализации качества «аномальности» (когда результат может быть субъективно воспринят как ошибочный). Предложенный подход позволяет объективизировать смысловую аномальность текстовых документов по отношению к заданной коллекции, благодаря чему снижается число ошибок и повышается точность решения.

На двух исследуемых коллекциях метод показал высокую эффективность, которая, однако, должна быть проверена в реальном применении в рамках прикладной информационной системы. Изучение откликов пользователей на предложенные системой оценки аномальности документов позволит подтвердить на практике точность представленного метода и его применимость на различных наборах данных. Составление соответствующих метрик для встраивания в систему является одним из возможных направлений дальнейших исследований.

Кроме того, для прикладного применения предложенного алгоритма целесообразно автоматизировать процесс определения порогового значения аномальности, который устанавливается через выбор документа, заведомо инородного для коллекции. Таким документом может являться внешний документ, который не соответствует коллекции тематически и стилистически. В рамках данной работы подбор такого инородного документа происходил вручную, на основе экспертной оценки. Без привлечения эксперта пороговое значение аномальности может устанавливаться на основе максимального значения скоринга

аномальности документов коллекции, с некоторым сдвигом этого значения, вычисленного при начальных параметрах алгоритма. Эта гипотеза также требует дальнейшего изучения и экспериментальной проверки.

Общее направление дальнейших исследований определяется движением в сторону т.н. машиночитаемого права и включает в себя решение задач, связанных с обнаружением смысловых аномалий в юридически значимых текстах: выявление новшеств в проектах российских нормативно-правовых документов и в локальных нормативных актах, формальное представление семантического содержания русскоязычных регулятивных документов [39], поиск отдельных аномальных структур в юридических текстах, обнаружение нормотворческих коллизий.

6. Заключение. В работе исследовалась задача обнаружения аномалий как документов, обладающих новизной по отношению к однородной текстовой коллекции, с учетом требований, который накладываются на выбор методов решения задачи прикладными интеллектуальными информационными системами (ПИИС), работающими с юридически значимыми документами. К этим требованиям относится высокая точность решения, реализуемая эффективность алгоритмов, воспроизводимость результата и его объяснимость.

Предложена стратегия выбора оптимального метода поиска аномалий и подбора его параметров в зависимости от предполагаемой в ПИИС векторной модели коллекции документов.

К коллекции добавляется заведомо инородный документ, с учетом которого определяются критерии оптимизации: максимальное различие распределений расстояний между документами коллекции до ее центра и до инородного документа; отсутствие (либо минимальное число) документов исходной коллекции, которые по значению индекса аномальности превосходят инородный документ. Точкой отсчета для индекса аномальности считается центр рассматриваемой коллекции.

Для точечного выявления аномальности необходимо определить в коллекции пограничный документ, имеющий максимальную величину индекса аномальности среди всех документов коллекции (функция определения аномальности выбирается в зависимости от векторной модели документов). В рамках ПИИС аномальность нового документа определяется в рекомендательном режиме, при этом может быть установлен допустимый лимит превышения индекса аномальности нового документа относительно пограничного документа.

Эксперимент проводился на двух однородных коллекциях юридически значимых русскоязычных документах: государственные стан-

дарты в сфере информационных технологий и в сфере железнодорожного сообщения. В качестве образцов инородных документов использовались отрывок из художественного произведения, статья в СМИ, научная статья, финансовая отчетность и шаблон договора.

Результаты эксперимента показали, что предложенный метод подбора алгоритмов и параметров по минимизации выбросов относительно заведомо инородного документа позволяет выбрать оптимальный алгоритм обнаружения аномалий для каждой текстовой коллекции, а полученная таким образом оценка аномальности документа соответствует экспертной оценке.

Для задач поиска аномалий в рамках процесса тематической кластеризации юридически значимых текстовых документов эффективен метод изолирующего леса (Isolation Forest), поскольку инородные документы в силу отсутствия явно выделенной темы изолируются быстрее, чем документы, принадлежащие семантически однородному кластеру.

Если для решения прикладной задачи используется разреженная полноразмерная векторная модель на основе словарной статистики, то после выбора оптимальных параметров словаря модели для ПИИС рекомендуется применить метод опорных векторов в модификации One-Class SVM с соответствующей функцией преобразования признакового пространства. При снижении размерности наилучший результат показывают линейные методы.

В ходе дальнейших исследований планируется протестировать предложенный метод на практике в рамках действующей прикладной информационной системы и оценить применимость метода через встроенные метрики, учитывающие отклики пользователей на предложенные системой оценки аномальности документов. Также целесообразно автоматизировать процесс определения порогового значения аномальности, используя данные о скоринге аномальности документов исходной коллекции. Другим направлением исследований станет доработка алгоритма поиска аномалий методом тематической кластеризации. В частности, требует проверки предположение, что использование словосочетаний вместо n-грамм при построении словаря модели улучшит определение аномальных тем.

В рамках дальнейших исследований решение актуальной прикладной задачи выявления смысловых аномалий в юридически значимых, в том числе регулятивных, документах позволит сделать шаг в сторону интеллектуализации нормотворческой деятельности и машиночитаемого права.

Литература

1. *Mahapatra A., Srivastava N., Srivastava J.* Contextual anomaly detection in text data // Algorithms. 2012. vol. 5. no. 4. pp. 469-489.
2. *Ghosal T. et al.* Novelty goes deep. A deep neural solution to document level novelty detection // Proceedings of the 27th International Conference on Computational Linguistics, 2018. pp. 2802–2813.
3. *Zhao L., Zhang M., Ma S.* The nature of novelty detection // Information Retrieval. 2006. vol. 9. no. 5. C. 521–541.
4. *Guzman J., Poblete B.* On-line relevant anomaly detection in the Twitter stream: an efficient bursty keyword detection model // Proceedings of the ACM SIGKDD workshop on outlier detection and description. 2013. pp. 31-39.
5. *Lau J. H. et al.* Word sense induction for novel sense detection // Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. 2012. pp. 591-601.
6. *Гурина А.О., Гузев О.Ю., Елисеев В.Л.* Обнаружение аномальных событий на хосте с использованием автокодировщика // International Journal of Open Information Technologies. 2020. Т. 8. №. 8.
7. *Goldstein M., Dengel A.* Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm // KI-2012: Poster and Demo Track. 2012. pp. 59-63.
8. *Zhao Y., Nasrullah Z., Li Z.* Pyod: A python toolbox for scalable outlier detection // arXiv preprint arXiv:1901.01588. 2019.
9. *Denning D.E.* An intrusion-detection model // IEEE Transactions on software engineering. 1987. no. 2. pp. 222-232.
10. *Markou M., Singh S.* Novelty detection: a review—part 1: statistical approaches // Signal processing. 2003. vol. 83. no. 12. pp. 2481-2497.
11. *Chandola V., Banerjee A., Kumar V.* Anomaly detection: A survey // ACM computing surveys (CSUR). 2009. vol. 41. no. 3. pp. 1-58.
12. *Pimentel M.A.F. et al.* A review of novelty detection // Signal Processing. 2014. vol. 99. pp. 215-249.
13. *Faria E.R. et al.* Novelty detection in data streams // Artificial Intelligence Review. 2016. vol. 45. no. 2. pp. 235-269.
14. *Ruff L. et al.* A unifying review of deep and shallow anomaly detection // Proceedings of the IEEE. 2021.
15. *Hendrycks D., Mazeika M., Dietterich T.* Deep anomaly detection with outlier exposure // arXiv preprint arXiv:1812.04606. 2018.
16. *Gorokhov O., Petrovskiy M., Mashechkin I.* Convolutional neural networks for unsupervised anomaly detection in text data // International Conference on Intelligent Data Engineering and Automated Learning. Springer, Cham, 2017. pp. 500-507.
17. *Yang Y. et al.* Topic-conditioned novelty detection // Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. 2002. pp. 688-693.
18. *Ng K.W. et al.* Novelty detection for text documents using named entity recognition // 2007 6th international conference on information, communications & signal processing. IEEE, 2007. pp. 1-5.
19. *Amplayo R.K., Hong S.L., Song M.* Network-based approach to detect novelty of scholarly literature // Information Sciences. 2018. vol. 422. pp. 542-557.

20. *Li Z. et al.* COPOD: copula-based outlier detection // arXiv preprint arXiv:2009.09463. 2020.
21. *Mikolov T., Yih W., Zweig G.* Linguistic regularities in continuous space word representations // Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies. 2013. pp. 746-751.
22. *Краснов Ф.В., Смазневич И.С.* Фактор объяснимости алгоритма в задачах поиска схожести текстовых документов // Вычислительные технологии. 2020. Т. 25. №. 5. С. 107-123.
23. *Schubert E., Gertz M.* Intrinsic t-stochastic neighbor embedding for visualization and outlier detection // International Conference on Similarity Search and Applications. Springer, Cham, 2017. pp. 188-203.
24. McInnes L., Healy J., Melville J. Umap: Uniform manifold approximation and projection for dimension reduction // arXiv preprint arXiv:1802.03426. 2018.
25. *Narayan A., Berger B., Cho H.* Density-preserving data visualization unveils dynamic patterns of single-cell transcriptomic variability // bioRxiv. 2020.
26. *Campos G.O. et al.* On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study // Data mining and knowledge discovery. 2016. vol. 30. №. 4. pp. 891-927.
27. *Amarbayasgalan T., Jargalsaikhan B., Ryu K.H.* Unsupervised novelty detection using deep autoencoders with density-based clustering // Applied Sciences. 2018. vol. 8. no. 9. pp. 1468.
28. *Campello R.J.G.B. et al.* Hierarchical density estimates for data clustering, visualization, and outlier detection // ACM Transactions on Knowledge Discovery from Data (TKDD). 2015. vol. 10. no. 1. pp. 1-51.
29. *Ankerst M. et al.* OPTICS: Ordering points to identify the clustering structure // ACM Sigmod record. 1999. vol. 28. no. 2. pp. 49-60.
30. *Karypis G., Han E.H., Kumar V.* Chameleon: Hierarchical clustering using dynamic modeling // Computer. 1999. vol. 32. no. 8. pp. 68-75.
31. *Karypis G., Kumar V.* A software package for partitioning unstructured graphs, partitioning meshes, and computing fill-reducing orderings of sparse matrices // University of Minnesota, Department of Computer Science and Engineering, Army HPC Research Center, Minneapolis, MN. 1998. vol. 38.
32. *Kannan R. et al.* Outlier detection for text data // Proceedings of the 2017 SIAM international conference on data mining. Society for Industrial and Applied Mathematics, 2017. pp. 489-497.
33. *Zhang J., Ghahramani Z., Yang Y.* A probabilistic model for online document clustering with application to novelty detection // Advances in neural information processing systems. 2004. vol. 17. pp. 1617-1624.
34. *Manevitz L. M., Yousef M.* One-class SVMs for document classification // Journal of machine Learning research. 2001. vol. 2. no. Dec. pp. 139-154.
35. *Zimek A., Campello R.J.G.B., Sander J.* Ensembles for unsupervised outlier detection: challenges and research questions a position paper // ACM SIGKDD Explorations Newsletter. 2014. vol. 15. no. 1. pp. 11-22.
36. *Marques H.O. et al.* Internal evaluation of unsupervised outlier detection // ACM Transactions on Knowledge Discovery from Data (TKDD). 2020. vol. 14. no. 4. pp. 1-42.
37. *Liu F.T., Ting K.M., Zhou Z.H.* Isolation Forest // 2008 Eighth IEEE international conference on data mining. IEEE, 2008. pp. 413-422.

38. *Краснов Ф.В.* Сравнительный анализ точности методов визуализации структуры коллекции текстов // International Journal of Open Information Technologies. 2021. Т. 9. №. 4. С. 79-84.
39. *Пименов В.И., Воронов М.В.* Формализация регулятивных текстов // Информатика и автоматизация. 2021. № 3 (20). С. 562–590.

Краснов Федор Владимирович — кандидат технических наук, эксперт, департамент семантических систем, NAUMEN R&D. Область научных интересов: интеллектуальная аналитика текстов. Число научных публикаций – 69. fkrasnov@naumen.ru, www.naumen.ru; 620028, Екатеринбург, ул. Татищева, 49А; р.т.: +7 (981)781-48-47.

Смазневич Ирина Сергеевна — бизнес-аналитик, департамент семантических систем, NAUMEN R&D. Область научных интересов: применение интеллектуальных алгоритмов в прикладных информационных системах. Число научных публикаций – 2. ismaznevich@naumen.ru, www.naumen.ru; 620028, Екатеринбург, ул. Татищева, 49А; р.т.:+7(916)722-66-39.

Баскакова Елена Николаевна — ведущий системный аналитик, департамент семантических систем, NAUMEN R&D. Область научных интересов: применение интеллектуальных алгоритмов в прикладных информационных системах. enbaskakova@naumen.ru, www.naumen.ru; 620028, Екатеринбург, ул. Татищева, 49А; р.т.: +7(903)258-75-93.

DOI 10.15622/ia.20.4.5

F. KRASNOV, I. SMAZVEVICH, E. BASKAKOVA

OPTIMIZATION APPROACH TO SELECTING METHODS OF DETECTING ANOMALIES IN HOMOGENEOUS TEXT COLLECTIONS

Krasnov F., Smaznevich I., Baskakova E. **Optimization Approach to Selecting Methods of Detecting Anomalies in Homogeneous Text Collections.**

Abstract. The problem of detecting anomalous documents in text collections is considered. The existing methods for detecting anomalies are not universal and do not show a stable result on different data sets. The accuracy of the results depends on the choice of parameters at each step of the problem solving algorithm process, and for different collections different sets of parameters are optimal. Not all of the existing algorithms for detecting anomalies work effectively with text data, which vector representation is characterized by high dimensionality with strong sparsity.

The problem of finding anomalies is considered in the following statement: it is necessary to checking a new document uploaded to an applied intelligent information system for congruence with a homogeneous collection of documents stored in it. In such systems that process legal documents the following limitations are imposed on the anomaly detection methods: high accuracy, computational efficiency, reproducibility of results and explicability of the solution. Methods satisfying these conditions are investigated.

The paper examines the possibility of evaluating text documents on the scale of anomaly by deliberately introducing a foreign document into the collection. A strategy for detecting novelty of the document in relation to the collection is proposed, which assumes a reasonable selection of methods and parameters. It is shown how the accuracy of the solution is affected by the choice of vectorization options, tokenization principles, dimensionality reduction methods and parameters of novelty detection algorithms.

The experiment was conducted on two homogeneous collections of documents containing technical norms: standards in the field of information technology and railways. The following approaches were used: calculation of the anomaly index as the Hellinger distance between the distributions of the remoteness of documents to the center of the collection and to the foreign document; optimization of the novelty detection algorithms depending on the methods of vectorization and dimensionality reduction. The vector space was constructed using the TF-IDF transformation and ARTM topic modeling. The following algorithms have been tested: Isolation Forest, Local Outlier Factor and One-Class SVM (based on Support Vector Machine).

The experiment confirmed the effectiveness of the proposed optimization strategy for determining the appropriate method for detecting anomalies for a given text collection. When searching for an anomaly in the context of topic clustering of legal documents, the Isolating Forest method is proved to be effective. When vectorizing documents using TF-IDF, it is advisable to choose the optimal dictionary parameters and use the One-Class SVM method with the corresponding feature space transformation function.

Keywords: Anomaly Detection, Novelty Detection, Outlier Detection, Homogeneous Text Collections, Sparse Space Dimension Reduction, Topic Modeling.

Krasnov Fedor — Ph.D., Expert, Department of Semantic Systems, NAUMEN R&D. Research interests: Intelligent text analysis. The number of publications – 69; e-mail: fkrasnov@naumen.ru, www.naumen.ru; 49A, Tatishcheva street, «Tatishchevsky» Business Center, 4th floor, Yekaterinburg, 620028, Russian Federation; office phone: +7 (981)781-48-47.

Smaznevich Irina — Business analyst, Department of Semantic Systems, NAUMEN R&D, Research interests: Use of intelligent algorithms in applied information systems. The number of publications – 2; e-mail: ismaznevich@naumen.ru, www.naumen.ru; 49A, Tatishcheva street, «Tatishchevsky» Business Center, 4th floor, Yekaterinburg, 620028, Russian Federation; office phone: +7(916)722-66-39.

Baskakova Elena — System analyst, Department of Semantic Systems, NAUMEN R&D. Research interests: Use of intelligent algorithms in applied information systems; e-mail: enbaskakova@naumen.ru, www.naumen.ru; 49A, Tatishcheva street, «Tatishchevsky» Business Center, 4th floor, Yekaterinburg, 620028, Russian Federation; office phone: +7(903)258-75-93.

References

1. Mahapatra A., Srivastava N., Srivastava J. Contextual anomaly detection in text data. *Algorithms*. 2012. vol. 5. no. 4. pp. 469-489.
2. Ghosal T. et al. Novelty goes deep. A deep neural solution to document level novelty detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, COLING 2018, Santa Fe, New Mexico, USA, August 20–26, 2018, pp. 2802–2813.
3. Zhao L., Zhang M., Ma S. The nature of novelty detection. *Information Retrieval*. 2006. vol. 9. no. 5. C. 521–541.
4. Guzman J., Poblete B. Online relevant anomaly detection in the Twitter stream: an efficient bursty keyword detection model. *Proceedings of the ACM SIGKDD workshop on outlier detection and description*. 2013. pp. 31-39.
5. Lau J. H. et al. Word sense induction for novel sense detection. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 2012. pp. 591-601.
6. Gurina A.O., Guzev O.Ju., Eliseev V.L. [Detection of anomalous events on the host using an autoencoder] Obnaruzhenie anomal'nyh sobytij na hoste s ispol'zovaniem avtokodirovshhika. *International Journal of Open Information Technologies*. 2020. vol. 8. no. 8.
7. Goldstein M., Dengel A. Histogram-based outlier score (HBOS): A fast unsupervised anomaly detection algorithm. *KI-2012: Poster and Demo Track*. 2012. pp. 59-63.
8. Zhao Y., Nasrullah Z., Li Z. Pyod: A python toolbox for scalable outlier detection. *arXiv preprint.* arXiv:1901.01588. 2019.
9. Denning D.E. An intrusion-detection model. *IEEE Transactions on software engineering.* 1987. no. 2. pp. 222-232.
10. Markou M., Singh S. Novelty detection: a review—part 1: statistical approaches. *Signal processing*. 2003. vol. 83. no. 12. pp. 2481-2497.
11. Chandola V., Banerjee A., Kumar V. Anomaly detection: A survey. *ACM computing surveys (CSUR).* 2009. vol. 41. no. 3. pp. 1-58.
12. Pimentel M.A.F. et al. A review of novelty detection. *Signal Processing*. 2014. vol. 99. pp. 215-249.
13. Faria E.R. et al. Novelty detection in data streams. *Artificial Intelligence Review*. 2016. vol. 45. no. 2. pp. 235-269.
14. Ruff L. et al. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE.* 2021.
15. Hendrycks D., Mazeika M., Dietterich T. Deep anomaly detection with outlier exposure. *arXiv preprint.* arXiv:1812.04606. 2018.
16. Gorokhov O., Petrovskiy M., Mashechkin I. Convolutional neural networks for unsupervised anomaly detection in text data. *International Conference on Intelligent Data Engineering and Automated Learning.* Springer, Cham, 2017. pp. 500-507.
17. Yang Y. et al. Topic-conditioned novelty detection. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2002. pp. 688-693.
18. Ng K.W. et al. Novelty detection for text documents using named entity recognition. *2007 6th international conference on information, communications and signal processing*. IEEE, 2007. pp. 1-5.

19. Amplayo R.K., Hong S.L., Song M. Network-based approach to detect novelty of scholarly literature. *Information Sciences*. 2018. vol. 422. pp. 542-557.
20. Li Z. et al. COPOD: copula-based outlier detection. *arXiv preprint.* arXiv:2009.09463. 2020.
21. Mikolov T., Yih W., Zweig G. *L*inguistic regularities in continuous space word representations. *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*. 2013. pp. 746-751.
22. Krasnov F.V., Smaznevich I.S. [The explicability factor of the algorithm in the problems of searching for the similarity of text documents]. *Vychislitel'nye tehnologii*. [Computational technologies]. 2020. vol. 25. no. 5. pp. 107-123.
23. Schubert E., Gertz M. *I*ntrinsic t-stochastic neighbor embedding for visualization and outlier detection. *International Conference on Similarity Search and Applications*. Springer, Cham, 2017. pp. 188-203.
24. McInnes L., Healy J., Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint.* arXiv:1802.03426. 2018
25. Narayan A., Berger B., Cho H. Density-preserving data visualization unveils dynamic patterns of single-cell transcriptomic variability. *bioRxiv*. 2020.
26. Campos G.O. et al. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data mining and knowledge discovery*. 2016. vol. 30. №. 4. pp. 891-927.
27. Amarbayasgalan T., Jargalsaikhan B., Ryu K.H. Unsupervised novelty detection using deep autoencoders with density-based clustering. *Applied Sciences*. 2018. vol. 8. no. 9. P. 1468.
28. Campello R.J.G.B. et al. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD).* 2015. vol. 10. no. 1. pp. 1-51.
29. Ankerst M. et al. OPTICS: Ordering points to identify the clustering structure. *ACM Sigmod record*. 1999. vol. 28. no. 2. pp. 49-60.
30. Karypis G., Han E. H., Kumar V. Chameleon: Hierarchical clustering using dynamic modeling. *Computer*. 1999. vol. 32. no. 8. pp. 68-75.
31. Karypis G., Kumar V. A software package for partitioning unstructured graphs, partitioning meshes, and computing fill-reducing orderings of sparse matrices. University of Minnesota, Department of Computer Science and Engineering, Army HPC Research Center, Minneapolis, MN. 1998. vol. 38.
32. Kannan R. et al. Outlier detection for text data. *Proceedings of the 2017 siam international conference on data mining. Society for Industrial and Applied Mathematics*, 2017. pp. 489-497.
33. Zhang J., Ghahramani Z., Yang Y. A probabilistic model for online document clustering with application to novelty detection. *Advances in neural information processing systems*. 2004. vol. 17. pp. 1617-1624.
34. Manevitz L. M., Yousef M. One-class SVMs for document classification. *Journal of machine Learning research*. 2001. vol. 2. no. Dec. pp. 139-154.
35. Zimek A., Campello R.J.G.B., Sander J. Ensembles for unsupervised outlier detection: challenges and research questions a position paper. *ACM SIGKDD Explorations Newsletter*. 2014. vol. 15. no. 1. pp. 11-22.
36. Marques H. O. et al. Internal evaluation of unsupervised outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*. 2020. vol. 14. no. 4. pp. 1-42.
37. Liu F.T., Ting K.M., Zhou Z.H., Isolation Forest. *2008 Eighth IEEE international conference on data mining.* IEEE, 2008. pp. 413-422.
38. Krasnov F.V. [Comparative Analysis of the Accuracy of Methods for Visualizing the Structure of a Text Collection]. *International Journal of Open Information Technologies.* 2021. vol. 9. no. 4. pp. 79-84. (In Russ.)
39. Pimenov V.I., Voronov M.V. [Formalization of regulatory texts]. *Informatika i avtomatizacija.* [Computer Science and Automation]. 2021. no. 3(20). pp. 562–590. (In Russ.)

DOI 10.15622/ia.20.4.6

N. MOUMOUTZIS, Y. SIFAKIS, S. CHRISTODOULAKIS,
D. PANEVA-MARINOVA, L. PAVLOVA

PERFORMATIVE FRAMEWORK AND CASE STUDY FOR TECHNOLOGY-ENHANCED LEARNING COMMUNITIES

Moumoutzis N., Sifakis Y., Christodoulakis S., Paneva-Marinova D., Pavlova L. **Performative Framework and Case Study for Technology-Enhanced Learning Communities.**

Abstract. This paper employs the overarching concept of communities to express the social contexts within which human creativity is exercised and learning happens. With the advent of digital technologies, these social contexts, the communities we engage in, change radically. The new landscape brought about by digital technologies is characterized by new qualities, new opportunities for action, new community affordances. The term onlife is adopted from the Onlife Manifesto and used to distinguish the new kind of communities brought about by the modern digital technologies, the onlife communities. Design principles are presented to foster such communities and support their members. These principles constitute a framework that emphasizes the concept of performativity, i.e. knowledge is based on human performance and actions done within certain social contexts, rather than development of conceptual representations. To demonstrate the use of the framework and the corresponding principles, the paper presents how they can be used to analyze, evaluate and reframe a concrete system addressing creativity and learning in the field of cultural heritage (history teaching and learning). One of the most significant results is the adoption of principles that facilitate students' engagement in rich learning experiences moving from the role of end-user towards the role of expert-user with the support of so called maieuta-designers. The result of this process is the use of the studied software not only to consume ready-made content but the creation of new, student generated content, offering new learning opportunities to the students. As the evaluation shows, these new learning opportunities enable students to develop a deeper understanding of the topics studied.

Keywords: Learning Communities, Performativity, Creativity, Evaluation.

1. Introduction. With the proliferation of modern computing technologies and the wide use of social networks to promote new personal learning opportunities, many people worldwide are engaged in technology-enhanced learning communities to pursue a personalized approach to technology use and learning while at the same time collaborate with other people pursuing common goals [1]. Official educational systems try to take advantage of these new opportunities and transcend traditional ways of teaching towards new approaches that put learners at the center of learning process and enable them to become creators of new content using appropriate tools and exploiting plenty of materials available online. This new way of learning is ideal for addressing any kind of learning theme ranging from official school curriculum subjects to life-long learning settings, especially when it comes to the need to continuously update and extend the knowledge of professionals and teachers in related disciplines [2].

Recognizing and promoting technology-enhanced learning within the above overall framework, there is a clear need to design and develop appropriate learning tools and platforms and use them within adequate learning and training frameworks. They offer diverse learning modes that are informed by modern pedagogical approaches that promote personalization and rich interactions and give opportunities for knowledge construction in a personally meaningful manner.

An overarching approach that can effectively address all these distinct learning settings can be based on the notion of community. A community is essentially a group of people who share an interest or have a common goal. The group can evolve naturally because of the members' common interest in a particular domain or area, or it can be created specifically with the goal of gaining knowledge related to the interests of its members. This knowledge development aspect can be effectively supported by an appropriate framework that provides an effective connection between learning and certain skills that can be demonstrated via specific performed tasks of the learner.

It is through the process of sharing information and experiences with the group that the members learn from each other, and have an opportunity to develop themselves. Communities can exist online, such as within discussion boards and newsgroups, or in real life, such as in a lunch room at work, at the institution or elsewhere in any social setting.

The goal of the work reported in this paper and its scientific novelty, following the discussion above, is to provide a comprehensive approach to technology-enhanced learning that is based on the concept of community as a central notion for enabling learning and creativity. The approach is informed by current trends in re-conceptualizing and rethinking about our societies facing the so called “hyperconnected era” [3]. This is reflected in the term “onlife”, which has been employed in The Onlife Manifesto [3]. This term stresses the fact that the deployment of information and communication technologies and their uptake by society radically affect the human condition, modifying our relationships to ourselves, to others and to the world. Consequently, the term onlife communities is employed to signify a social context that emerges from these developments.

To elaborate a framework for the establishment and support of onlife communities, valuable lessons can be drawn from work reported in [4, 5]. In particular, that work criticizes the so called engineering mythology that is based on a set of certain assumptions. These assumptions are related to the fact that digital systems consist of parts that interact according to certain patterns that are defined and understood at design time. Designers seek to discover these patterns, or even invent them, with the aim to achieve a certain way of operation of the system under consideration. On the other hand, users

are expected to interact with the final system in certain ways following certain rules that are effectively imposed by the internal logic of the system.

Following the engineering mythology in technological projects there are cases when the emergence of certain patterns in end-user usage of the systems or desirable features that go beyond initial assumptions during to address these emergent patterns, an alternative mythology has been proposed that goes beyond the legitimacy of design as a process done by computing experts with the participation of representatives of end-users in order to guide the development of digital systems [4]. This alternative mythology points to the fact that digital systems can be realized by the composition of elementary components with limited initial design and be put to work by end-users, facilitated by computing experts that play the role of catalysts of change and evolution of those systems towards directions not initially fore-seen [4].

Elaborating on this alternative design approach, this paper presents PerFECt, a Performative Framework to Establish and Sustain Onlife Communities, i.e. communities of creators using digital tools in a certain domain, emphasizing creativity and learning. This framework is then employed to analyze and evaluate the use of ViSTPro, a spatiotemporal process visualization platform. Spatiotemporal processes are a generic model for representing various types of content knowledge [6, 7, 8]. Such processes are difficult for learners to conceptualize with traditional teaching and learning approaches due to their complexity and inherent dynamic nature [9]. Consequently, a generic approach on dynamic spatiotemporal process modeling could be used in supporting many educational domains, promoting an active inquiry-based style to learning [9].

The ViSTPro platform presents features that can be analyzed via the PerFECt framework concepts: it fosters active explorations of spatiotemporal processes as rich scenarios prepared by educators and offered to learners through a web-based player. Playing a scenario involves graphic representation of formations, movements, and interactions on Google Maps. This way, learners interact with graphical entities in an intuitive manner. In contrast, traditional ways of learning depend upon a painful and difficult process to develop abstract mental images with no real-world direct mapping. Additionally, the platform enables learners to make questions and receive personalized explanations. Therefore, learners can watch the representation of the processes' evolution in space and time, and actively intervene. Furthermore, ViSTPro can offer new learning opportunities to leaners if they are enabled to use the functionality initially offered to teachers (scenario authoring). This means, the students can be supported to act as creators of their own scenarios, thus becoming expert users, in the terminology of the PerFECt framework. This is an option that clearly demonstrates the applicability of the concepts

of the framework to foster learning communities and facilitating a transformation process for end-users to become expert users.

The rest of the paper is organized as follows: Section 2 presents the details of the PerFECt framework along with links to the related literature that provides important concepts used in the framework. Section 3 presents the core features of the ViSTPro platform and links ViSTPro to the PerFECt framework by presenting its use in order to interpret how users apply or could apply ViSTPro. Section 4 presents experimental evaluation results on the usage of ViSTPro in the domain of learning communities (learning his-tory) that confirm the hypotheses drawn from applying the PerFECt framework. Section 5 concludes and presents directions for future work.

2. The PerFECt Framework and Background Work. As exposed in [5, 10], end-users of digital systems are increasingly more required to act as active contributors at use time, thus becoming "producers" of contents and functionalities. The term expert-user is to signify a person that is an expert in a particular domain with main goal to develop the capabilities of available software tools. An expert-user subsumes all those roles denoting people in charge of carrying out creative/authoring activities without being a professional software developer. Usually, the role of end-user and that of an expert-user is played by different people that may also belong to different communities. Furthermore, [5, 10] suggest the role of meta-designer to describe professionals who create the socio-technical conditions for empowering expert-users to engage in continuous system development. Meta-designers create open systems at design time that can evolve by their users acting as co-designers. Yet another important role is that of maieuta-designer who is mainly oriented at organizational and social issues, rather than technical ones, for supporting the task of the expert-users: ensuring the socio-technical prerequisites that are necessary for enabling expert-users working out new solutions by using the available technological means. This task undertaken by expert-users addresses as many end-users as possible in the process of continuous refinement of the available technology, thus promoting and strengthening participation, as the ultimate goal of maieuta-designers. The word "maieuta" is used in direct analogy to the well know learning method employed by Socrates, the philosopher. It signifies the facilitation of people to address challenges by enabling them to develop knowledge and self-confidence and ultimately transform themselves from passive consumers of technology to active creators, i.e. moving from the role of end-user towards the role of expert-user.

Starting from the above conceptualization of user roles of meta-designers, maieuta-designers, end-users and expert-users, the PerFECt framework seeks to adapt these concepts within the so called hyper connected context that is framed by modern digital technologies. This is captured in the

term onlife that is borrowed from the Onlife Manifesto [3] to describe the type of communities that this framework is trying to describe and establish. The Onlife Manifesto is the result of work within the Onlife Initiative that started as a project envisioned and implemented directly by the European Commission's Information Society Directorate-General in 2012. The project was intended to explore the extent to which the digital transition impacts societal expectations towards policy making. The project outcome was the Onlife Manifesto [3]. The baseline of this text is that advent of digital technologies in all aspects of life has fundamental consequences in the human condition. It affects our reference frameworks, in a number of different domains including:

- our self-conception (who we are);
- our mutual interactions (how we socialize);
- our conception of reality (our metaphysics);
- our interactions with reality (our agency).

The members of the Onlife Initiative decided to adopt the neologism "onlife" to refer to the new experience of a hyperconnected reality within which it is no longer sensible to ask whether one may be online or offline. Within this new reality that is brought about by digital technologies and their ever-increasing pervasiveness four important transformations are happening:

- the blurring of the distinction between reality and virtuality;
- the blurring of the distinctions between human, machine and nature;
- the reversal from information scarcity to information abundance;
- the shift from the primacy of entities to the primacy of interactions

Following these developments, the PerFECt framework suggests the term onlife community to signify aggregations that emerge in hyperconnected spaces when humans engage with other humans as well as with machines and natural entities in mindful interactions with sufficient human feeling to form webs of personal relationships. Furthermore, by adopting the four user roles of end-user, expert-user, meta-designer and maieuta-designer it seeks to provide a certain structure to onlife communities and provide a mechanism to enable rich learning experiences.

To further analyze how these user roles are understood in their dynamics, it is important to note that they interact with each other and with digital artifacts and digital tools to form a co-evolution phenomenon. The meta-designer is focused on designing and providing the most effective tools that may sustain the co-evolution between end-users and expert-users. The maieuta-designer facilitates the migration from the role of end-user to the role of expert-user to empower end-users to appropriate and contribute to the use of available digital tools. In cases when an end-user is not interested or fails to

evolve into the role of expert-user, the maieuta-designer may facilitate participation in system evolution by systematizing the reporting of shortcomings and system faults as identified by the end-user and proposing solutions that are handled by expert-users.

Consequently, the above four roles give rise to two co-evolution processes: The first one refers to the use of software targeted to the end-user where there is continuous (cyclical) interaction between the end-user and the system. This is depicted in Figure 1 (left) with three homocentric cycles of arrows that represent the action-interpretation cycle at the lower level, the task-object cycle at the middle level and community-technology cycle at the upper level. In an analogous way, there is a second cyclical process depicted in Figure 1 (right) that refers to the use of software components as building blocks of the system in continuous evolution from the perspective of expert-users. This process corresponds to yet another three homocentric cycles of the same nature: action-interpretation, task-object, and community-technology layers.

The inner interaction cycle in each co-evolution process refers to actions (triggered by the corresponding user or software) that are interpreted by the other party (software or user respectively). The task-object cycle in the middle refers to the co-evolution of the user task and the corresponding artifact within the boundaries of the system. Finally, an outer community-technology cycle captures the idea that the overall environment within which a user is working (community), co-evolves with the technology that supports the operation of the environment.

Before describing in more detail the two co-evolution processes, one pertaining to end-users and the other to expert-users, we need to present the concept of universality. This concept refers to blends of machines and physical objects that generalize the notion of software or tool within a hyperconnected landscape. Universality addresses the issue of causality in digital representations, as Brenda Laurel puts it in her seminal book “Computers as Theatre”: “The fact that people seek to understand causality in representational worlds provides the basis for Aristotle’s definition of universality. In the colloquial view, an action is universal if everybody can understand it, regardless of cultural and other differences among individuals. This would seem to limit the set of universal actions to things that everyone on the planet does: eat, sleep, love, etc. Aristotle posits that any action can be “universalized” simply by revealing its cause; that is, under-standing the cause is sufficient for understanding the action, even if it is something alien to one’s culture, back-ground, or personal ‘reality’.” [11]. Consequently, a Universal Object is an artifact that presents itself in a way that is meaningful and understandable through casual relationships that enable the user of such an object

to effectively manipulate it and understand it, i.e. link it to casual interpretations. A Universal Object can be acted upon to produce certain effects because its casual interpretation enables the user to know what will happen if certain manipulations are made. Furthermore, its response to certain manipulations is predictable and thus can be used to produce the desired effects within the context used. The essence of digital technologies is, in this respect, to transform plain objects into Universal Objects.

Universal Objects are considered a core element of the PerFECt framework and their use by end-users in combination with their development (as Universalizing Assemblies) by expert-users constitute the two aforementioned co-evolution phenomena in three co-eccentric cycles built around them to describe the relationship between end-users/expert-users and the end-tasks/expert-tasks that they engage in. This phenomenon has been first described in [12] and has been linked to an approach to effectively address end-user needs during system design and evolution. End-user needs evolve as the end-users use a specific technology meaning that the system developers need to support the evolution of the systems as well as to adapt and address the evolving end-user needs. In a similar way, expert-users' needs evolve as well.

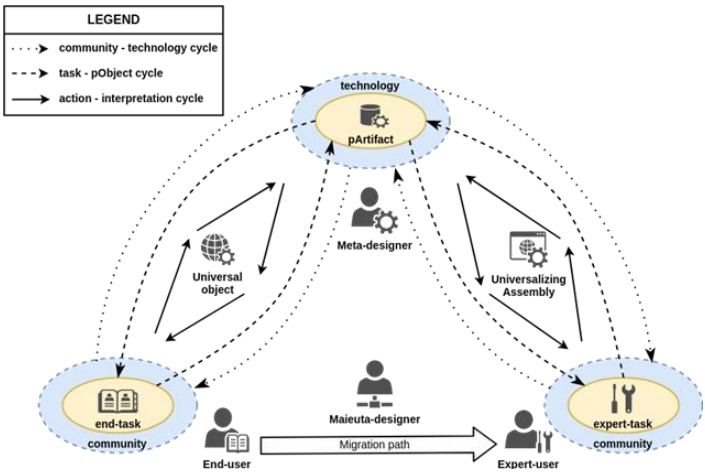


Fig. 1. Main components of the PerFECt framework

Let us see in more detail the co-evolution phenomenon around Universal Objects as depicted in the left part of Figure 1:

1. At the higher level, there is the community-technology of co-evolution cycle meaning. Relationship of people and technology is dynamic and evolves people working with a certain technology to learn how to do

certain things. As a result of this evolution their expectations and conceptualizations with respect to available technology change and give rise to evolution of the technology itself that triggers further evolution of the community of end-users. This higher level of co-evolution cycle entails the motives of people, i.e. their needs and how they satisfy their needs via their activities. By offering new interaction possibilities, technologies employed change end-user habits and this means that the social and work organization evolves with the use of certain technologies.

2. At a lower level, there is the task-artifact co-evolution cycle that refers to the tasks that end-users are able to do with a specific version of a system and the corresponding artifacts that they manipulate or use during their work. Consequently, at this level, end-users articulate their behavior towards certain goals that form a cause-effect chain in order to pursue the motives of the upper level. Furthermore, the use of certain artifacts to support end-users' tasks suggest in many cases new possible tasks and these new tasks mean that new artifacts should be created.
3. At the lowest level there is an interaction cycle during which end-users are expected to do certain operations to effectively use the available technological features. Such interactions call for a certain interpretation of their actions in order to be able to effectively use the available features. This lowest level cycle could thus be conceptualized by successive materialization and interpretation, representing meaningful actions that support and trigger the upper levels of the co-evolution phenomenon.

At the center of the three end-user cycles (left part of Fig. 1) the PerFECt framework suggests the concept of Universal Object, a concept that generalizes the concept of software and it is based on the concept of universality as already presented. Such Universal Objects, can be the result of the work of expert-users, as will be described next, to address the evolving needs of end-users within the wider context provided by the PerFECt framework.

The co-evolution cycles that address the expert-user role (right part of Fig. 1) are structured around the concept of Universalizing Assembly. This is a complementary concept to the concept of Universal Object. A Universalizing Assembly is essentially a synthesis of performative artifacts (pArti-facts) that enables the creation of Universal Objects supporting the task of end-users. Consequently, the task of expert-users is to enable this universalization of plain objects by exploiting the available tools in the form of performative artifacts (pArtifacts) to account for the incorporation of the idea of performativity in digital technologies.

Performativity underlines the relationship between humans and the artifacts they create that is triggered by social interaction and continuously recreates the bonds that keep the society as a whole. [13] offers an interesting concept to capture this idea and link to purposeful and mindful use of physical objects: the concept of performative object, which is a special type of design object to facilitate mindful awareness of the physical and symbolic social actions and their consequences. Considering that performative objects are design objects, the framework presented here uses the term performative artifacts in a broader sense: all artifacts involve a certain level of performativity that is usually captured by their affordances i.e. clues about how an object should be used, typically provided by the object itself or its context. However, this latter term, does not explicitly refer to mindful-ness as a target during the design process. In this respect, the term performative artifact, is used here to capture the idea of intentional design for social interaction, to create and sustain social bonds and call for symbolic social actions that recreate the social contexts within which we live in. To conclude, the three levels of the co-evolution cycles in the case of the expert-users, are the following:

1. At the higher level, the community-technology co-evolution cycle captures the dynamics of the relationship of people and technology using pArtifacts to enable the universalization of certain objects. As expert-users evolve and learn how to do certain things (i.e. providing tools in the form of Universal Objects to end-users) their expectations and conceptualizations with respect to the available technology change and give rise to evolution of the technology it-self that triggers further evolution of the community as a whole.
2. At a lower level, there is the task-artifact co-evolution cycle that refers to the tasks that expert-users are able to do with a specific version of a system and the corresponding pArtifacts that they use to create/extend the Universalizing Assemblies that constitute the target of their work. It is important to note that expert-users pro-vide the ground for end-users to work within a cause-effect framework (this is captured by the idea of universalization). The evolution of expert-users is related to the evolution of end-users as well taking into account that the role of expert-users is to support the evolving needs of end-users. To this end, there is a critical contribution of maieuta-designers that facilitate the articulation of end-user needs and their effective communication to the expert-users.
3. At the lowest level there is an interaction cycle during which ex-pert-users are expected to do certain operations when they are engaged in the use of the technological features offered by meta-designers. Such interactions call for a certain interpretations so that expert-users are

able to use the underlying technologies in an effective way. At this level, the interaction between the expert-users and the underlying technologies are taken without direct reference to their social context as successive rounds of materialization and interpretation of certain actions that support and trigger the upper levels of the co-evolution phenomenon.

To summarize, an onlife community within the PerFECt framework captures the structure that is imposed by four user roles of the framework: end-user, expert-user, maieuta-designer and meta-designer along with the artifacts, tools and even underlying physical objects to account for situations where technologies are embedded into an underlying reality. In other words, the adoption of the concept of community emphasizes the fact that all these user roles, through their interactions within the two co-evolution processes, create a bigger aggregation of humans. They engage with other humans as well as with machines and natural entities in mindful interactions, thus creating the social contexts described as onlife communities to account for hyperconnectivity as well.

3. A Case Study: ViSTPro. As a concrete case of study, the framework presented in the previous section can be put in action to enable a deeper understanding of how modern digital technologies can foster social interaction and promote creativity and learning. This section presents ViSTPro and how it enables the formation of onlife communities within the domain of cultural heritage and, in particular, history learning.

ViSTPro has been developed by the authors of this paper as a generic tool to enable the visualization of spatiotemporal processes. Such processes are a generic model for representing various types of content knowledge ranging from historical developments (e.g. representations of battles and other historical events) to physical processes like the ones studied in geosciences, life sciences etc. [6, 7, 8]. Such processes are difficult for learners to conceptualize with traditional teaching and learning approaches due to their complexity and inherent dynamic nature [9]. Consequently, a generic approach on dynamic spatiotemporal process modeling could be used in supporting many educational domains, promoting an active inquiry-based style to learning.

To address this need, interactive digital maps could be employed, on top of which appropriate active objects are overlaid representing real-world phenomena or events. This way, open exploration environments could be assembled and offered to learners as intuitive dynamic spatiotemporal models [9]. The learner has the opportunity to create symbolic formations, move them, and determine their interaction with other entities. In other words, a

"visual narrative" is developed, in such a setting, for the evolution of spatio-temporal processes trough the animation of imported, process-type specific, graphical symbols superimposed on maps. Learners are familiar with this kind of animations because they regularly use digital applications with similar features [14]. The transfer of this positive experience in a learning context brings pleasure and offers more opportunities for learner engagement.

To effectively support this type of learning and engage learners, careful design should be employed. ViSTPro employs the concept of scenario for modeling complex spatiotemporal processes [15]. This concept suggests the visual representation of the evolution of spatiotemporal processes. Exploitations and semantic maps play an important role in this representation. ViSTPro distinguishes scenario authoring from scenario playback. During scenario authoring ViSTPro helps and guides the scenario author through-out the process. The scenario author initially selects a name and describes the new scenario. At the same time, active components of the scenario are determined.

A scenario contains groupings, types of entities and specific entities. For example, in the case of a historical battle a grouping may represent the troops that participate in the battle, types of entities may relate to the infantry or cavalry and certain types of entities can represent leading figures of the battle. The user selects the characteristic color of each troop and the representation of each entity type and also can import additional icons. Specific types of entities are represented with a larger size in order to differentiate from other types of entities. In addition, the representation of the different states of the types of entities providing multiple views for each of them (e.g. killed, on-fire, etc.) can be supported, while the user can create custom states. These elements are contained in optional map legends to facilitate the explanatory power of the presentation.

The second stage of the authoring process is the structuring of the scenario. The structural elements of a scenario include activities, sub-activities and events. Activities correspond to main units of action. Each activity is connected with a title, a description, and may include other activities or elementary unities of action (sub-activities) where the action unfolds and the movements, actions and interactions of the active components are visually described. For each subactivity several properties are available such as its name, description, start- and end-time, photos, recorded narration and related activities and other sub-activities of the scenario. Sub-activities may include events that represent a milestone or a particular incident. An event is identified by its title, description, timestamp and possibly its correlation with some type of entities, its state and a semantic object. Each scenario is thus modeled as a hierarchical structure of activities, sub-activities and events.

Another important scenario element is the set of formations that will be visualized. A formation is a set of entities handled as a whole. ViSTPro offers the necessary tools for the design of formations through predefined geometrical shapes (square, rectangle, circle, polygon, etc.), varying sizes, orientations, etc. After formation definition, entity types can be specified along with their size, location and density, in order to be included in the corresponding formations. Furthermore, the existence and position of one or more specific types of entities can be indicated.

The handling of a specific formation is possible through sub-activities. When a sub-activity is created, the scenario author chooses which formations will appear, defines the initial and final position and specifies the path that will be followed during scenario playback. Furthermore, there is a set of actions available for each formation. These actions are related to their behavior and interaction during playback.

The representation of actions is displayed by means of suitable graphical elements, such as icons and arrows. A formation may change its state as it moves or performs an action or interacts with other formations during scenario playback. For this reason, during scenario authoring, it is possible to redefine the state of a formation by defining its size, shape and density of varying types of entities, while their state can be modified.

Another important modeling primitive is graphics. A set of graphical elements are available such as lines, arrows, and other predefined elements, which are overlaid on the map during scenario authoring, and they play a crucial role in the playback of a scenario. Semantic content is provided via title and description, and can also determine characteristics such as color, size and orientation. During scenario playback the graphics can remain stationary or move. They can also change their shape in a manner similar to the state change of formations.

Process visualization addresses important elements such as human-made objects and significant locations of the surroundings. The presentation and provision of relevant information regarding these objects are done through semantic maps. Semantic maps are collections of important locations and objects of a region, which are represented on a map. The creation of a semantic map gives the possibility to create semantic objects each one described by its name, description, and one or more images. Thus, during scenario playback, it is possible to interact with the objects of a semantic map and examine their semantic content. Semantic maps are customizable by selecting certain objects and creating a new semantic map containing them. The new semantic map can be saved with a new name for future use. An original or customized semantic map can be used in one or more scenarios in the way described above.

During scenario playback individual learning needs are addressed through the provision of explanations for better understanding the evolution of the processes represented in each scenario. ViSTPro handles the movement of formations, involved in each sub-activity from an initial to a final position and provides an intuitive representation of state changes by changing the size, shape and density and status of the types of entities employing appropriate interpolations. Furthermore, during scenario playback each sub-activity title and description is presented possibly enriched with sound recorded narration. The playback can be paused to give time for examination of photos, related information that may have been registered in the sub-activity or even the physical surroundings. Events are depicted through entitled panels on the map, with location and time properly indicated. If an event is associated with a specific type of entities and/or a specific semantic object, those entities and/or objects are shown emphasized. Event-related additional information and pictures can be examined if scenario playback is paused.

During the playback of a scenario its hierarchical structure is provided. Through this structure one can switch to another scenario that describes in more detail the currently presented sub-activity. Finally, it is possible to speed up or slow down playback in order to adjust the speed to learner needs.

For more details regarding the design of the ViSTPro and its comparison with similar software targeting spatiotemporal process visualization, see [15]. The objective here is to present how ViSTPro use in specific learning settings can be guided by the PerFECt framework. In other words, to present a comprehensive case study that demonstrates how to employ the concepts and user roles of the PerFECt framework presented in Section 2. Analyze the use of software platforms such as ViSTPro so that they can be put within a wider context that accounts for the rich social interactions that could be promoted towards the establishment of onlife communities. In particular, ViSTPro can be considered as a representative tool on how a learning community can be established (in the field of cultural heritage in general and history learning in particular) that brings together:

- software developers supporting the software and providing further enhancements to address the needs of the users,
- teachers that prepare animations of historical events, i.e. scenarios representing the corresponding spatiotemporal processes in ViSTPro along with semantic maps and digital materials explaining the details of the animated events, and
- students that use the scenarios prepared by teachers to learn about the animated historical events in a personalized manner.

Employing the user roles described by the PerFECt framework, the above categories of participants in a ViSTPro-based learning community can be presented as follows:

Software developers that support ViSTPro and implement further enhancements to address the needs of teachers and students are the meta-designers of the PerFECt framework. As meta-designers, they are expected to be offered an open system that can evolve by its users as co-designers. To enable this, ViSTPro offers several capabilities to use various media types, thus offering the capability to integrate digital materials coming from a diverse range of sources. Furthermore, it offers a flexible authoring environment as a means to support expert-users that wish to develop new scenarios, thus animating new historical events or providing alternative visualization for events that have been already described with existing scenarios. Finally, ViSTPro is also open with respect to creating semantic maps, i.e. providing semantic information about human-made objects and physical formations on top of Google maps. Semantic maps can be used within scenario playback to provide important semantic information that allow for deeper understanding of the animated events.

Scenario authors (e.g. teachers, but also historians or even students that wish to engage in activities to apply their historical knowledge in developing ViSTPro scenarios). They create scenarios in ViSTPro what the PerFECt framework describes as expert-users that address the needs of end-users using the open system capabilities offered by meta-designers to develop new components in the form of universalizing assemblies of digital objects that can then be used by end-users in their learning tasks. This is indeed what teachers are expected to do with ViSTPro: using its features to develop scenarios for spatiotemporal processes that represent important historical events. These scenarios capture knowledge about the corresponding historical events along with pedagogical content knowledge so that effective scaffolding can take place that will enable students to develop their historical knowledge within a rich learning environment supporting social interactions and use of digital tools to make complex historical events more understandable.

Students that use ViSTPro to see the animations of the scenarios (using its playback features) are what the PerFECt framework calls end-users. They essentially use the creations of expert-users in the form of Universal Objects, i.e. digital artifacts that represent and present causality within and across historical events to make the historical knowledge more understandable and justifiable.

Apart from the above-mentioned roles, which are directly related to ViSTPro as a tool supporting authoring and playback of spatiotemporal pro-

cesses, the PerFECt framework introduces yet another (fourth) user role: maieuta-designers. This is an important role that has a critical contribution in framing and supporting an onlife community. In particular, as already discussed, maieuta-designers are addressing the social conditions for supporting the task of expert-users and the transition from the end-user role to the role of expert-user. This transition and support of expert-users' tasks are essentially a learning process that takes place within a social context (i.e. the community of users). In the case of ViSTPro, the need for maieuta-designers emerges very naturally from the use of the tool in actual learning situations when students are enabled to develop their own scenarios (i.e. go beyond the end-user role toward the expert-user role) and thus learn deeper about the historical events they study. These enhanced learning results are documented by the actual evaluation following a controlled experiment approach in actual school settings presented in the next section.

Furthermore, teachers also express their belief that students can learn better when they are engaged in expert-user role tasks, thus creating their own portfolio of digital artifacts that can help them express their creativity and offer insights and motivation to find more information about the studied historical events using digital resources. Consequently, this approach is directly related to inquiry-based learning approaches and, more importantly, to constructionism: the learning theory that claims learners study better when they construct things [16].

4. Evaluation results. Teaching spatiotemporal processes, as it is the case for history courses in school, can be quite difficult and challenging if traditional approaches are employed. This is due to the fact that teaching of these subjects forces the student to develop complex cognitive structures. The task of correlating spatiotemporal information presented linearly when using traditional means is left to the student to handle without any significant help. E.g., the description of a battle in a history textbook presents the related events in a certain sequence although, in many cases, these events may happen in parallel and trigger other events in complex causal relationships. Consequently, traditional teaching results at construction of a vague cognitive structure, thus, complicating absorbing the knowledge of complex spatiotemporal processes. Using ViSTPro in teaching and learning of such processes can offer significant advantages and the purpose.

At accounting for the above issues, the purpose of ViSTPro evaluation was to study the performance of the system to facilitate understanding of spatiotemporal processes for students. Specifically, the evaluation addressed the system's use by teachers and students in teaching and learning of historical events matching the complex spatiotemporal processes. The aim was to

investigate whether ViSTPro promotes the effective learning at that also offering engagement and opportunities for creative expression. Thus, the investigation addressed both the use of ready-made scenarios and the creation of new scenarios by the students to demonstrate their knowledge and exercise their creativity. Furthermore, the evaluation aimed at studying the usability of ViSTPro and the emotional response of students when using it. Students evaluated the usability of the platform in terms of using the existing scenarios and interacting with them, as well as after experiencing the process of creating their own scenarios. The emotions evoked by the use of the system were also explored, as they shape the users' experience and influence upon the users' attitude towards ViSTPro.

The evaluation of ViSTPro was carried out in lower secondary schools (students aged 12-13) within an important topic of the history course: "The Battle of Marathon". In Greece the students of the 1st grade of high school study this important historical event. The experiment involved two classes of students, one serving as the control group and the other as the experimental group. Both classes were uniform in terms of the students' general performance. The first class (control group) consisted of 22 students (13 boys and 9 girls). The second class (experimental group) consisted of 17 students (12 boys and 5 girls). The students were clearly explained the purpose of the experiment and asked to answer anonymously through the used questionnaires.

The teachers who participated in the experiment also were clearly informed about the experiment purpose. The history teacher of the experimental group was trained to use ViSTPro in order to use it during the experimentation. The experimental group was also supported by the computer science teacher of the school since the experimental use of ViSTPro was exercised in the school computer lab. Both teachers and students out of the experimental group created their accounts in ViSTPro, thus, initiating a community that used the tool to share scenarios and learn through them as well as by creating new scenarios and sharing them with the other members of the community. The control group was offered a traditional teaching of the selected topic without any digital tool.

The experimental process was split in two stages. The purpose of the first stage was to study the effect of using ViSTPro when the teaching is performed based on ready-made scenarios. The experimental group was taught the selected topic using ViSTPro. The teaching was structured with due account for the functionality of the platform so that the features were used appropriately. Specifically, the teacher presented the battle and interacted with the presented scenario, activating and interrupting playback, and utilizing

system explanations. The students played an active role throughout the learning. Then, each student had an opportunity to interact with the scenario individually through a separate computer. The students in the control group were taught the same unit traditionally, covering the same body of knowledge and covering the same aspects of the selected topic. At the end of the first stage, students from both the control and the experimental group were asked to answer content questions, while those in the experimental group filled out additional questionnaires to give feedback on the usability of the platform, their emotional response to its use and a final one addressing the creativity aspects of using ViSTPro.

The second stage of the experimental process aimed at studying the effect of ViSTPro upon learning when students undertake the role of scenario creators. Therefore, only the students of the experimental group were engaged in this stage. The students constructed their own scenarios to represent the "Battle of Marathon". The construction of the students' scenarios additionally required six (6) teaching hours. Initially the students created their accounts in ViSTPro and were informed by their computer science teacher about the script creation process. The second lesson dealt with scenario creation, specification of groupings, and the creation of the necessary types of entities. The remaining teaching hours were used by the students to represent the battle stages. After accomplishing this process, the students again answered content questions on the selected topic and filled out the questionnaires on the usability and emotional response.

The content questionnaire used both for the control and the experimental group to measure their performance after the teaching of the selected topic totally consisted of 27 content questions: 20 multiple choice questions, 2 questions to arrange the facts in correct chronological order, 1 matching question and 4 free text questions. The completion time for the content questionnaire comprised 40 minutes.

The questions were categorized following the classification of Bloom's educational objectives [17]. The corresponding types of questions were:

- Knowledge (8 questions) – recalling specific facts and data.
- Comprehension (5 questions) – understanding what is taught without necessarily connecting with other materials or seeing full implications.
- Application (5 questions) – generalizing and using abstract information in specific situations.
- Analysis (4 questions in total) – splitting a problem/issue into subdivisions and detecting relations between them.
- Synthesis (3 questions) – assembling parts to form a whole.

– Evaluation (2 questions) – using criteria to make judgments.

The results drawn from the content questionnaires are shown below.

Table 1 compares the results of the control group and the experimental group at the first stage of the evaluation. Table 2 presents the results of the experimental group comparing between the first and the second stage of the evaluation. The grading of the student test used was based on a 100 scale.

Table 1. Results of the first stage of evaluation

	Experimental Group			Control Group		
	Mean	Stand. Error	95% T. Confid. Interval	Mean	Stand. Error	95% T. Confid. Interval
Knowledge	94.85	3.23	6.84	42.05	4.39	9.14
Comprehension	95.19	2.35	4.99	40.91	4.00	8.32
Application	77.78	5.87	12.45	26.26	4.11	8.54
Analysis	80.88	3.41	7.23	42.05	4.76	9.91
Synthesis	69.12	4.17	8.83	16.48	3.33	6.92
Evaluation	55.29	7.77	16.47	20.00	5.74	11.93

The results of the first evaluation stage (Table 1 and Figures 2, 3) clearly show the higher performance of the experimental group that used ViSTPro for all six categories of questions. It is worth noting that most of the averages of the experimental group exceeded 80/100, while those of the control group are under 50/100. The lower grades in the categories "Synthesis" and "Evaluation" are justified by the complexity of questions in these categories, in relation to the students' intellectual maturity. Nevertheless, the performance of the students of the experimental group is much higher disregarding the categories.

Figures 2 and 3 present the box plots of the performance of two groups so that the results can be seen in finer detail. The mean and median value of each question category is represented by dashed and bold lines respectively. The minimum and maximum performance scores of each category are also presented. The box plots confirm the results shown in Table 1: The performance of the students of the experimental group is much higher in all cases while the performance of the students in the control group is much lower. In addition, the last two categories of questions that are more demanding, demonstrate the importance of using ViSTPro in learning: The performance of the students in the experimental group in most cases exceeds 50/100 with respect to these two categories ("Synthesis" and "Evaluation"). There exist

few exceptions of very low scores, which could be considered as outliers and are presented as small circles on the graph. Therefore, it can be safely concluded that the ViSTPro use facilitates learning and helps students to reach in-depth understanding and achievement of demanding learning objectives.

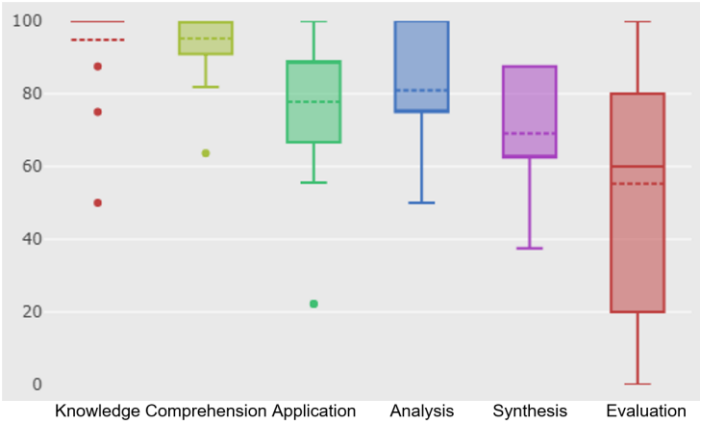


Fig. 2. Box plot of student performance (experimental group – first stage)

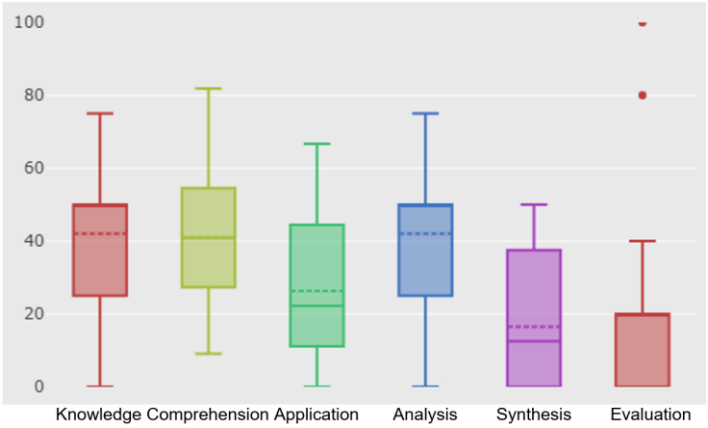


Fig. 3. Box plot of student performance (control group – first stage)

The experimental data out of the second stage of the evaluation present a further improvement in the performance of the students of the experimental group when they have used ViSTPro to create their own scenarios for the Battle of Marathon. Obviously, the improvement magnitude is small,

since the students already had, from the first stage, a rather high performance. However, the creation of scenarios by the students bore a qualitative improvement that concerned both the learning process and the acceptance of the platform as students realized its potential to enhance their learning.

The students got excited, gave better structured answers and got acquainted with the technology. Table 2, below, summarizes the results and Figure 4 presents the corresponding box plot.

Table 2. Results of the first and second evaluation stages for the experimental group

	First Stage			Second Stage		
	Mean	Stand. Error	95% T. Confid. Interval	Mean	Stand. Error	95% T. Confid. Interval
Knowledge	94.85	3.23	6.84	97.79	2.21	4.68
Comprehension	95.19	2.35	4.99	95.19	2.71	5.75
Application	77.78	5.87	12.45	87.58	5.46	11.58
Analysis	80.88	3.41	7.23	80.88	4.03	8.54
Synthesis	69.12	4.17	8.83	70.59	4.54	9.62
Evaluation	55.29	7.77	16.47	72.94	6.85	14.52

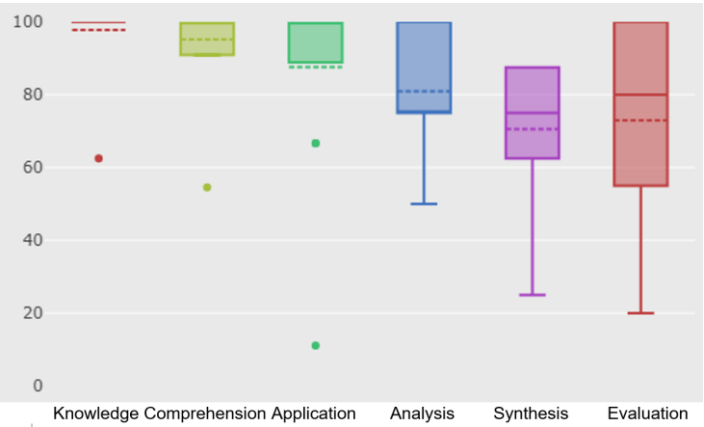


Fig. 4. Box plot of student performance (experimental group – second stage)

Beyond the learning effectiveness of ViSTPro presented above, another important aspect of the evaluation was the usability of ViSTPro. This part of the evaluation was made using the System Usability Scale (SUS) questionnaire [18]. SUS is a general-purpose questionnaire and can be used in

evaluating the usability of different systems, equipment and products. It consists of 10 questions (statements) graded by the respondents on a 5-point scale which is numbered from 1 ("Strongly Disagree") to 5 ("Strongly Agree"). The final score measures the usability of the system under evaluation.

The usability evaluation was done by the students of the experimental group at both the first and the second stage of the evaluation. At the first stage the final value of the SUS questionnaire was equal to 75.88, a value that characterizes the system as acceptable to good. At the second stage, the students rated the system with a final SUS value of 78.53. This shows that the demanding process of creating scenarios was followed without any significant difficulties and proves for the case as well the usability of the system. Besides that, the second stage showed that the students dealt with the system with pleasure, enthusiasm and good mood. Especially for the aspect of learnability (measuring how easily the users can learn a system), whose value is derived from two particular questions of SUS, the results are very important. The students graded the learnability of ViSTPro at the first stage by 78.68 and at the second stage by 83.82.

The diagram in Figure 5 shows the SUS score during the first stage (blue line) and the second stage (green line) of the evaluation on top of the benchmark graph used to interpret SUS scores as described in [18].

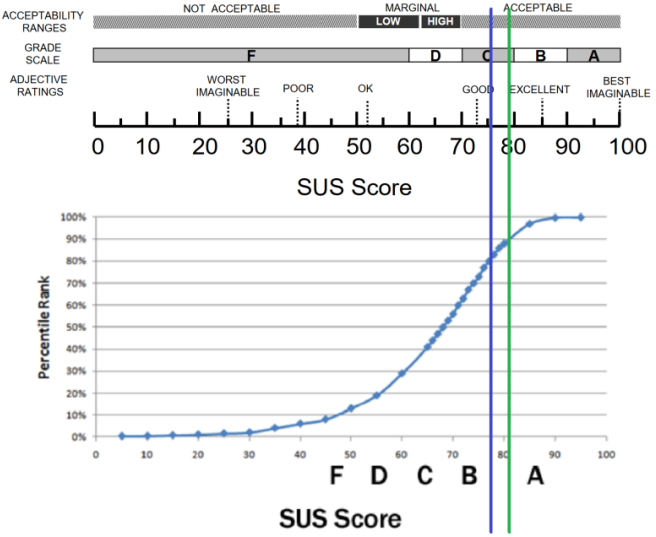


Fig. 5. SUS score during the first stage (blue line) and the second stage (green line) of the evaluation

In close relation to the usability, the emotional response of students towards ViSTPro was also evaluated. The methodology adopted for the evaluation of the emotional response uses emoticon cards [19]. These cards contain sixteen faces, male and female, depicting distinct emotions. These figures are grouped in pairs, each representing a combination of two emotional states. Therefore, the cards can be divided into eight sectors: Calm-Pleasant, Calm-Neutral, Calm-Unpleasant, Average-Unpleasant, Excited-Unpleasant, Excited-Neutral, Excited-Pleasant, and Average-Pleasant. The scores that fall into the pleasant sectors are interpreted as positive.

The following two figures (Fig. 6, 7) depict the results of this part of the evaluation. Both figures refer to the participants of the experimental group, during the evaluation first and second stages respectively.

The majority of choices in both cases is on the right side of the chart what means that the students found it pleasant to engage with ViSTPro. It should be noted that the results improved significantly at the second stage of the experiment, where the number of students that were excited using the system doubled. This result is in harmony with usability evaluation results and demonstrates that while the simple use of the system (first stage) provides for excellent learning opportunities, students find the creation of scenarios (second stage) more exciting. Therefore, using ViSTPro by students for the scenarios creating is an effective learning tool that is highly acceptable by them, promotes their creativity and ensures their active participation in the learning process.

Regarding creativity, the corresponding evaluation tool used was a short questionnaire based on the analysis presented in [20]. In particular, in the above work the authors suggest the use of a series of questions and metrics that examine whether the system under consideration helps to develop creativity. In resonance with their suggestions, the questionnaires used had the following questions:

- Q1: Would you use ViSTPro for personal reasons?
- Q2: Would you use ViSTPro to share a scenario with your friends?
- Q3: Would you use ViSTPro to spread a message to many people you do not know (e.g., online)?
- Q4: Would you use ViSTPro to share something you learned with your classmates?
- Q5: Would you like to watch scenarios made with ViSTPro?
- Q6: Do you think that ViSTPro could be used for educational purposes?

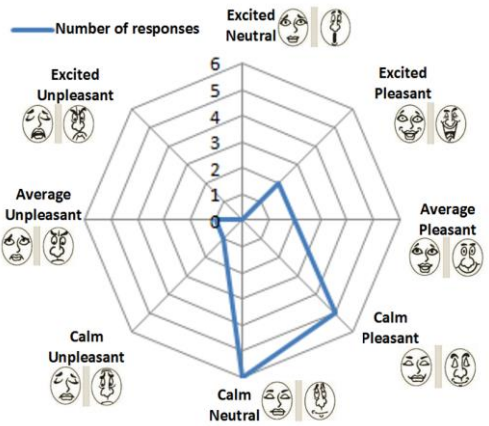


Fig. 6. Emotional response results (first stage)

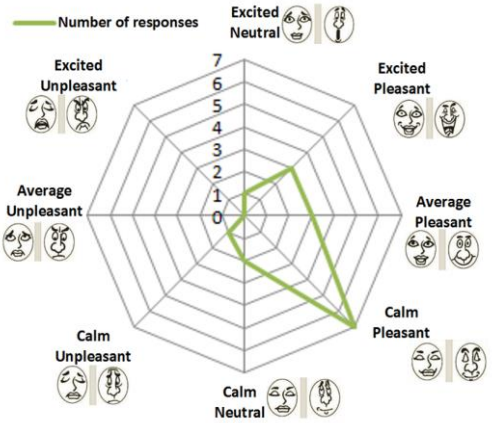


Fig. 7. Emotional response results (second stage)

Figure 8 shows the box plot results for the questions above. Students would use the system for the reasons they were asked, as the average of positive answers is over 3.7 in all cases. The median value exceeds (except for one question) the average and consistently has the value 4. The acceptance of the system by the students as an educational tool is impressive, since the average of the grades is equal to 4.2 and the median with 5. Therefore, students would use ViSTPro for personal purposes, to communicate their social environment and to follow scenarios and consider it a tool with significant educational value.

To summarize the results presented, using ViSTPro in teaching and learning with respect to traditional approaches has a significant positive impact. This is documented by the learning outcomes observed between the control group and the experimental group during the first stage of the evaluation. Apart from this, the results of the second stage of the evaluation that refer to students undertaking the role of scenario authors, demonstrate how the concepts of the PerFECt framework can be put in action and foster interactions to promote creativity and engagement in learning. In particular, the second stage of the evaluation addresses all three co-evolution processes depicted in Figure 1 that are in the core of the PerFECt framework. The action-interpretation cycles, referring to the way end-users and expert-users interact with the underlying system, are captured by the usability evaluation done using the SUS questionnaire. The task-oriented cycles are captured by the emotional response evaluation. Finally, the community-technology cycles, are captured by the creativity questionnaire and the content questionnaire evaluating the learning effectiveness of ViSTPro.

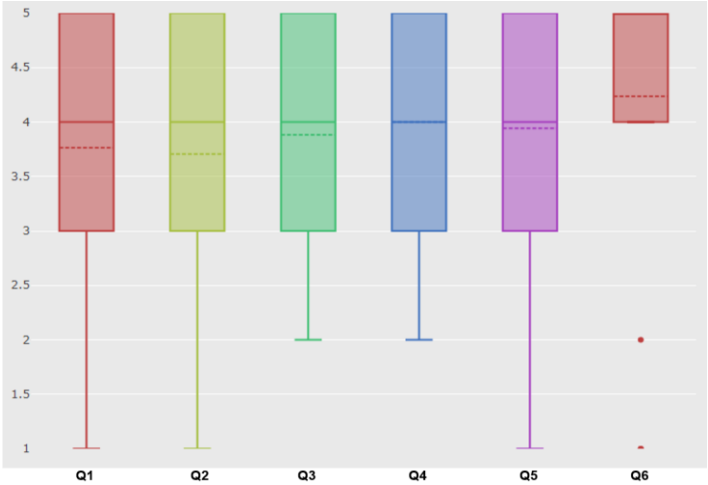


Fig. 8. Box plot with result for creativity evaluation

In all cases, the positive effects of the students undertaking the role of expert-users (i.e. scenario creators) are clearly shown. To enable this positive effect, the presence of maieuta-designers that will facilitate students in their transition from end-users to expert-users is extremely important. Moreover, it is important to underline that ViSTPro effectively supports the formation of a community of users by providing all the necessary features to create and manage user accounts, create and share scenarios and semantic maps.

6. Conclusion. The PerFECt framework presented in this paper addresses issues related to the establishment and support of onlife communities as rich socio-technical contexts where learning can be promoted, and engaging learning experiences can take place. This framework can be used to inform the design and use of educational digital platforms and tools.

As a concrete example of such a tool and its use, this paper presents ViSTPro, a digital platform that enables the specification of scenarios used to animate spatiotemporal processes. ViSTPro puts heavy emphasis in the hierarchical structure of spatiotemporal processes and enables gradual study from higher level abstractions to more specific detailed representations. Thus, a learner may initially see a simplified version of process evolution and then go into more details. The movements, actions and interactions of the formations are graphically represented on maps and those belonging to subactivities are presented synchronized based on a common timeline. This way the learner becomes familiar with complex spatiotemporal processes and acquires a complete picture of their development in time.

To demonstrate the flexibility of the ViSTPro and its capability to provide insightful visualizations, a pilot application has been elaborated focusing on the Battle of Marathon, pertaining to the corresponding chapter in the History course of the first grade of Greek Gymnasium (grade K-7). Based on this application, a controlled experiment was designed to evaluate the learning effectiveness of history teaching by ViSTPro. The evaluation also addressed ViSTPro usability, emotional response of users and creativity support. The results clearly demonstrate that using ViSTPro, history teaching becomes more engaging, and the learning outcomes are much better. An important part of the evaluation was to offer students the possibility to create their own scenarios for the taught topic. This additional task resonates with the PerFECt framework concept of facilitating the migration from end-user to expert-user role. The participating students were thus engaged in more deep study of the subject. An improvement of their performance was observed as well as higher scores in usability evaluation of ViSTPro and their students' emotional response to it. Furthermore, the students were extremely enthusiastic about the use of the tool and reported that it was much more interesting and engaging for them to create their own version of the scenario to study the Battle of Marathon instead of just watching the playback of a readymade scenario.

Future work will further employ the PerFECt framework to facilitate the use of ViSTPro in other learning domains to support users with alternative objectives and address new needs beyond history learning within school curriculum. Such objectives could be related to the study of local history by members of local communities that gradually evolve to expert-users and de-

velop their own scenarios to represent the current state of knowledge regarding certain events of the local history. In such scenarios, the supported features of ViSTPro can be put under new light to account for supporting new semantics thus opening up new capabilities for the users.

Future work will further explore the use of the PerFECt framework to better understand how other software applications and systems promoting creativity and learning could be enhanced and repurposed to promote rich social interactions. Such a platform that will be analyzed and enhanced in the proposed manner is eShadow [21]. eShadow promotes an innovative digital storytelling approach inspired by traditional shadow theatre and it also provides extensions addressing other storytelling traditions such as digital marionettes. The analysis of eShadow using the PerFECt framework will identify interesting workflows that address the transition of its users from the end-user role to the expert-user role, how related external tools can promote this transition and facilitate the development of digital media authoring skills. Yet another domain that will be addressed refers to learning personalization [22, 23] addressing issues related to the use of digital tools to offer learning opportunities tailored to personal needs and expectations by individual users. Finally, a very interesting domain to analyze using the lens of the PerFECt framework is serious games [24-26] taking into account its importance in developing student creativity, learning engagement and collaboration within communities.

References

1. Linton J.N. Institutional factors for supporting electronic learning communities. *Online Learning*. 2017. 21(1). pp. 238-256.
2. Brown B.D., Horn R. S., King G. The effective implementation of professional learning communities. *Alabama Journal of Educational Leadership*. 2018. 5. pp. 53-59.
3. Ganascia J.G. Views and Examples on Hyper-Connectivity. In: Floridi, L. (ed.), *The Onlife Manifesto: Being Human in a Hyperconnected Era*. Springer. 2015.
4. Cabitza F., Simone C. Building socially embedded technologies: Implications about design. In *Designing Socially Embedded Technologies in the Real-world*. Springer London. 2015. pp. 217-270.
5. Fischer G., Fogli D., Piccinno A., Revisiting and broadening the meta-design framework for end-user development. In *New perspectives in end-user development*. Springer, Cham. 2017. pp. 61-97.

6. Pant N., Fouladgar M., Elmasri R., Jitkajornwanich K. A Survey of Spatio-Temporal Database Research. In Asian Conference on Intelligent Information and Database Systems. *Springer,* Cham. 2018. pp. 115-126.
7. Siabato W. An Annotated Bibliography on Spatio-temporal Modelling Trends. *International Journal of Earth & Environmental Sciences*. 2017.
8. Firat E.E., Laramee R.S. Towards a survey of interactive visualization for education. EG UK Computer Graphics & Visual Computing, Eurographics Proceedings. 2018.
9. Resnick M., Turtles, termites, and traffic jams: Explorations in massively parallel microworlds. *MIT Press*. 1997.
10. Cabitza F., Fogli D., Piccinno A. Cultivating a Culture of Participation for the Co-Evolution of Users and Systems. In *CoPDA@ AVI*. 2014. pp. 1-6.
11. Laurel, B., Computers as theatre. Addison-Wesley. 2013.
12. Fogli D., Piccinno. A Co-evolution of end-user developers and systems in multi-tiered proxy design problems. In International Symposium on End User Development. Springer, Berlin, Heidelberg, 2013.
13. Niedderer K. Designing mindful interaction: the category of performative object. *Design issues*. 2007. 23(1). pp. 3-17.
14. Lameras P., Arnab S., Dunwell I., Stewart C., Clarke S., Petridis P. Essential features of serious games design in higher education: Linking learning attributes to game mechanics. *British Journal of Educational Technology.* 2017. 48(4). pp. 972-994.
15. Sifakis Y., Arapi P., Moumoutzis N., Christodoulakis S. ViSTPro: Spatiotemporal processes visualization in engineering education and crisis training. In IEEE Global Engineering Education Conference (EDUCON). 2017. pp. 413-422.
16. Kafai Y.B. Constructionist visions: Hard fun with serious games. *International Journal of Child-Computer Interaction*. 2018. 18, 19-21.
17. Ramirez T.V. On pedagogy of personality assessment: application of Bloom's Taxonomy of Educational Objectives. *Journal of Personality Assessment*. 2017. 99(2). pp. 146-152.
18. Brooke J. SUS: a retrospective. *Journal of usability studies*. 2013. 8(2). pp. 29-40.
19. Agarwal A., Meyer A. Beyond usability: evaluating emotional response as an integral part of the user experience. In CHI'09 Extended Abstracts on Human Factors in Computing Systems. 2009. pp. 2919-2930.

20. Hewett T., Czerwinski M., Terry M. Creativity support tool evaluation methods and metrics. *Creativity Support*. 2005. pp. 10-24.
21. Moumoutzis N., Christoulakis M., Christodoulakis S., Paneva-Marinova D. Renovating the Cultural Heritage of Traditional Shadow Theatre with eShadow. Design, Implementation, Evaluation and Use in Formal and Informal Learning. *Digital Presentation and Preservation of Cultural and Scientific Heritage*. Vol. 8, Sofia, Bulgaria: Institute of Mathematics and Informatics – BAS, 2018, pp. 51-70, ISSN 1314-4006 (Print), eISSN 2535-0366 (Online).
22. Yoshinov R.D., Iliev O.P. The structural way for binding a learning material with personal preferences of learners. *Proceedings SPIIRAS*. 2018. vol. 5, no. 60, pp. 189-215.
23. Arapi P. Supporting Personalized Learning Experiences on top of Multimedia Digital Libraries. Sofia: A dissertation submitted for the award of educational and scientific degree" Doctor of Philosophy. Bulgarian Academy of Sciences. 2017.
24. Márkus Z.L., Kaposi G., Veres M., Weisz Z., Szántó G., Szkaliczki T., Paneva-Marinova D., Pavolv R., Lunchev D., Goynov M., Pavlova L. Interactive game development to assist cultural heritage. In *Digital Presentation and Preservation of Cultural and Scientific Heritage*. vol. 8, Sofia, Bulgaria: Institute of Mathematics and Informatics – BAS. 2018. pp. 71-82, eISSN 2535-0366.
25. Paneva-Marinova D., Pavlov R. Mini-symposium on future trends in serious games for cultural heritage. *Digital Presentation and Preservation of Cultural and Scientific Heritage*. vol. 8, Sofia, Bulgaria: Institute of Mathematics and Informatics – BAS. 2018. pp. 241-244, eISSN 2535-0366.
26. Zhonggen Y. A meta-analysis of use of serious games in education over a decade. *International Journal of Computer Games Technology*. 2019.

Moumoutzis Nektarios — M.Eng., Ph.D., Researcher, Laboratory of Distributed Multimedia Information Systems and Applications, School of Electrical and Computer Engineering, Technical University of Crete, Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences. Research interests: eLearning systems and applications, cultural digital systems and applications, digital libraries, multimedia content management systems, semantic interoperability between digital libraries and eLearning systems, STEAM teaching and learning. The number of publications — 68. nektar@ced.tuc.gr, http://www.music.tuc.gr/Person.show?ID=10; Building

of Sciences, ECE School, University Campus, Kounoupidiana, Chania, 73100, Greece; office phone: +302821037395, fax: +302821037567.

Sifakis Yiannis — M.Eng., Junior researcher, Laboratory of Distributed Multimedia Information Systems and Applications, School of Electrical and Computer Engineering at the Technical University of Crete. Research interests: visualization software and applications, eLearning systems and applications, web engineering, STEM teaching and learning. The number of publications — 4. i_sifakis@hotmail.com; Building of Sciences, ECE School, University Campus, Kounoupidiana, Chania, 73100, Greece; office phone: +302821037395, fax: +302821037567.

Christodoulakis Stavros — Ph.D., Professor emeritus, Founder and former head of laboratory, Laboratory of Distributed Multimedia Information Systems and Applications, School of Electrical and Computer Engineering, Technical University of Crete. Research interests: databases, information systems, culture and tourism applications, digital libraries, e-learning infrastructures, high performance multimedia architectures, medical applications, multimedia content management systems, personalized and interactive TV applications, semantic interoperability, semantic natural language processing and interfaces, semantic web services, architectures and systems, semantics and personalization in multimedia services, semantics in 3D graphics and spatial applications, web application development methodologies. The number of publications — 272. stavros@ced.tuc.gr, http://www.music.tuc.gr/Faculty.show?ID=2; Building of Sciences, ECE School, University Campus, Kounoupidiana, Chania, 73100, Greece; office phone: +302821037399, fax: +302821037567.

Paneva-Marinova Desislava — Ph.D., Associate professor, Head of department, Mathematical Linguistics Department, Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences (BAS). Research interests: computer science, technologies for knowledge presentation and processing, data management and processing, data analytics, intelligent data curation, Semantic web, digital content management systems, web services. The number of publications — 112. dessi@cc.bas.bg; 8, Akad. G. Bonchev Str., 1113, Sofia, Republic of Bulgaria; office phone: +359888894814.

Pavlova Lilia — Ph.D., Assistant professor, Laboratory of Telematics, Bulgarian Academy of Sciences (BAS). Research interests: computer science, technologies for knowledge presentation and processing, Semantic web, e-learning. The number of publications — 34. pavlova.lilia@gmail.com; 8,

Akad. G. Bonchev Str., 1113, Sofia, Republic of Bulgaria; office phone: +35929793831.

Acknowledgements. This research is partly supported by the Bulgarian Ministry of Education and Science under the National Research Programme "Cultural heritage, national memory and development of society" and the National Scientific Program “Information and Communication Technologies for a Single Digital Market in Science, Education and Security” approved by DCM №577/17.08.2018. ViSTPro has been used in pilot activities with schools and professional associations in Chania, Greece, within the context of the EVANDE project (contract number ECHO/SUB/2014/693261) and the Erasmus+ projects DISCOVER (2017-1-BG01-KA202-036327), and MUSILIB (2018-1-FI01-KA201-047196).

УДК 004.8 **DOI** 10.15622/ia.20.4.6

Н. Мамуцис, С. Сифакис, С. Христодулакис, Д. Панева-Маринова, Л. Павлова

ПЕРФОРМАТИВНАЯ ПЛАТФОРМА И ЕЕ ПРИМЕНЕНИЕ ДЛЯ ВЫСОКОТЕХНОЛОГИЧНОГО ОБРАЗОВАТЕЛЬНОГО СООБЩЕСТВА

Мамуцис Н., Сифакис С., Христодулакис С., Панева-Маринова Д., Павлова Л. **Перформативная платформа и ее применение для высокотехнологичного образовательного сообщества.**

Аннотация. В этой статье используется всеохватывающая концепция сообществ для выражения социальных контекстов, в которых осуществляется человеческое творчество и происходит обучение. С появлением цифровых технологий эти социальные контексты, сообщества, в которых мы задействованы, радикально меняются. Новый ландшафт, созданный цифровыми технологиями, характеризуется новыми качествами, новыми возможностями для действий сообществ. Термин onlife заимствован из Манифеста Onlife и используется для обозначения сообществ нового типа, созданных современными цифровыми технологиями - сообществ onlife.

Представлены принципы проектирования, направленные на развитие таких сообществ и поддержку их членов. Эти принципы составляют основу, которая подчеркивает концепцию перформативности, то есть то, что знания основаны на деятельности человека и действиях, выполняемых в определенных социальных контекстах, а не на развитии концептуальных представлений. Чтобы продемонстрировать использование структуры и соответствующих принципов, в статье представлено, как их можно использовать для анализа, оценки и переформулирования конкретной системы, относя ее к творчеству и обучению в области культурного наследия (преподавание и изучение истории).

Одним из наиболее значительных результатов является принятие принципов, которые облегчают вовлечение студентов в учебный процесс, переходя от роли конечного пользователя к роли эксперта-пользователя при поддержке так называемых maieuta-дизайнеров. Результатом этого процесса является использование изученного программного обеспечения не только для потребления готового контента, но и для создания нового, сгенерированного студентами контента, предлагающего студентам новые возможности для обучения. Как показывает оценка, эти новые возможности обучения позволяют студентам развивать более глубокое понимание изучаемых тем.

Ключевые слова: учебные сообщества, перформативность, творчество, оценивание.

Мамуцис Нектариос — M.Eng., Ph.D., лаборатория распределенных мультимедийных информационных систем и приложений, Болгарская академия наук, Школа электротехники и вычислительной техники, Технический университет Крита, Институт математики и информатики. Область научных интересов: системы и приложения электронного обучения, культурные цифровые системы и приложения, электронные библиотеки, системы управления мультимедийным контентом, семантическая совместимость между цифровыми библиотеками и системами электронного обучения, преподавание и обучение в STEAM. Число

научных публикаций — 68. nektar@ced.tuc.gr, http://www.music.tuc.gr/Person.show?ID=10; Здание наук, школа ECE, университетский городок, Кунупидиана, Ханья, 73100, Греция; р. т.: +302821037395, факс: +302821037567.

Сифакис Яннис — M.Eng., младший научный сотрудник, лаборатория распределенных мультимедийных информационных систем и приложений, школа электротехники и вычислительной техники, Технический университет Крита. Область научных интересов: программное обеспечение и приложения для визуализации, системы и приложения электронного обучения, веб-инженерия, преподавание и обучение STEM. Число научных публикаций — 4. i_sifakis@hotmail.com; Здание наук, школа ECE, университетский городок, Кунупидиана, Ханья, 73100, Греция; р. т.: +302821037395, факс: +302821037567.

Христодулакис Ставрос — Ph.D., почетный профессор, основатель и руководитель лаборатории, лаборатория распределенных мультимедийных информационных систем и приложений, Школа электротехники и вычислительной техники, Технический университет Крита. Научные интересы: базы данных, информационные системы, приложения для культуры и туризма, цифровые библиотеки, инфраструктуры электронного обучения, высокопроизводительные мультимедийные архитектуры, медицинские приложения, системы управления мультимедийным контентом, персонализированные и интерактивные телевизионные приложения, семантическая совместимость, семантическая обработка естественного языка и интерфейсы, семантические веб-сервисы. , архитектуры и системы, семантика и персонализация в мультимедийных сервисах, семантика в трехмерной графике и пространственных приложениях, методологии разработки веб-приложений. Число научных публикаций — 272. stavros@ced.tuc.gr, http://www.music.tuc.gr/Faculty.show?ID=2; Здание наук, школа ECE, университетский городок, Кунупидиана, Ханья, 73100, Греция; р. т.: +302821037399, факс: +302821037567.

Панева-Маринова Десислава — Ph.D., доцент, заведующая кафедрой, кафедра математической лингвистики, Институт математики и информатики, Болгарская академии наук (БАН). Научные интересы: информатика, технологии представления и обработки знаний, управление и обработка данных, аналитика данных, интеллектуальное курирование данных, семантическая сеть, системы управления цифровым контентом, веб-сервисы. Число научных публикаций — 112. dessi@cc.bas.bg;

ул. Акад. Георги Бончев, бл. 8, 1113, София, Республика Болгария; р.т.: +359888894814.

Павлова Лилия — Ph.D., научный сотрудник, лаборатория телематики, Болгарская академии наук (БАН). Сфера научных интересов: информатика, технологии представления и обработки знаний, семантическая паутина, электронное обучение. Число научных публикаций — 34. pavlova.lilia@gmail.com; ул. Акад. Георги Бончев, бл. 8, 1113, София, Республика Болгария; р.т.: +35929793831.

Поддержка исследований. Исследование выполнено при частичной финансовой поддержке Министерством образования и науки Болгарии в рамках Национальной исследовательской программы «Культурное наследие, национальная память и развитие общества» и Национальная научная программа «Информационные и коммуникационные технологии для единого цифрового рынка в науке, образовании и безопасности», DCM № 577/17.08.2018. ViSTPro использовался в пилотных проектах со школами и профессиональными ассоциациями в Ханье, Греция, в контексте проекта EVANDE (номер контракта ECHO / SUB / 2014/693261) и проектов Erasmus + DISCOVER (2017-1-BG01-KA202-036327) и MUSILIB (2018-1-FI01-KA201-047196).

Литература

1. *Linton J.N.* Institutional factors for supporting electronic learning communities // Online Learning. 2017. 21(1). pp. 238-256.
2. *Brown B.D., Horn R.S., King G.* The effective implementation of professional learning communities // Alabama Journal of Educational Leadership. 2018. 5. pp. 53-59.
3. *Ganascia J.G.* Views and Examples on Hyper-Connectivity. Floridi, L. (ed.), The Onlife Manifesto: Being Human in a Hyperconnected Era. Springer. 2015.
4. *Cabitza F., Simone C.* Building socially embedded technologies: Implications about design // Designing Socially Embedded Technologies in the Real-world. Springer London. 2015. pp. 217-270.
5. *Fischer G., Fogli D., Piccinno A.* Revisiting and broadening the meta-design framework for end-user development // New perspectives in end-user development. Springer, Cham. 2017. pp. 61-97.
6. *Pant N., Fouladgar M., Elmasri R., Jitkajornwanich K.* A Survey of Spatio-Temporal Database Research // Proceedings of Asian Conference on Intelligent Information and Database Systems. Springer, Cham. 2018. pp. 115-126.

7. *Siabato W.* An Annotated Bibliography on Spatio-temporal Modelling Trends // International Journal of Earth & Environmental Sciences. 2017.
8. *Firat E.E., Laramee R.S.* Towards a survey of interactive visualization for education // EG UK Computer Graphics & Visual Computing, Eurographics Proceedings. 2018.
9. *Resnick M.* Turtles, termites, and traffic jams: Explorations in massively parallel microworlds. MIT Press. 1997.
10. *Cabitza F., Fogli D., Piccinno A.* Cultivating a Culture of Participation for the Co-Evolution of Users and Systems // CoPDA@ AVI. 2014. pp. 1-6.
11. *Laurel B.* Computers as theatre. Addison-Wesley. 2013.
12. *Fogli D., Piccinno* A. Co-evolution of end-user developers and systems in multi-tiered proxy design problems // International Symposium on End User Development. Springer, Berlin, Heidelberg, 2013.
13. *Niedderer K.* Designing mindful interaction: the category of performative object // Design issues. 2007. 23(1). pp. 3-17.
14. *Lameras P., Arnab S., Dunwell I., Stewart C., Clarke S., Petridis P.* Essential features of serious games design in higher education: Linking learning attributes to game mechanics // British Journal of Educational Technology. 2017. 48(4). pp. 972-994.
15. *Sifakis Y., Arapi P., Moumoutzis N., Christodoulakis S.* ViSTPro: Spatiotemporal processes visualization in engineering education and crisis training // IEEE Global Engineering Education Conference (EDUCON). 2017. pp. 413-422.
16. *Kafai Y.B.* Constructionist visions: Hard fun with serious games // International Journal of Child-Computer Interaction. 2018. 18, pp. 19-21.
17. *Ramirez T.V.* On pedagogy of personality assessment: application of Bloom's Taxonomy of Educational Objectives // Journal of Personality Assessment. 2017. 99(2). pp. 146-152.
18. *Brooke J.* SUS: a retrospective // Journal of usability studies. 2013. 8(2). pp. 29-40.
19. *Agarwal A., Meyer A.* Beyond usability: evaluating emotional response as an integral part of the user experience // CHI'09 Extended Abstracts on Human Factors in Computing Systems. 2009. pp. 2919-2930.
20. *Hewett T., Czerwinski M., Terry M.* Creativity support tool evaluation methods and metrics // Creativity Support. 2005. pp. 10-24.
21. *Moumoutzis N., Christoulakis M., Christodoulakis S., Paneva-Marinova D.* Renovating the Cultural Heritage of Traditional Shadow Theatre with eShadow. Design, Implementation, Evaluation and Use in

Formal and Informal Learning // Digital Presentation and Preservation of Cultural and Scientific Heritage. Vol. 8, Sofia, Bulgaria: Institute of Mathematics and Informatics – BAS, 2018, pp. 51-70.

22. *Yoshinov R.D., Iliev O.P.* The structural way for binding a learning material with personal preferences of learners // SPIIRAS Proceedings. 2018. vol. 5, no. 60, pp. 189-215.
23. *Arapi P.* Supporting Personalized Learning Experiences on top of Multimedia Digital Libraries // A dissertation submitted for the award of educational and scientific degree. Doctor of Philosophy. Bulgarian Academy of Sciences. 2017.
24. *Márkus Z.L., Kaposi G., Veres M., Weisz Z., Szántó G., Szkaliczki T., Paneva-Marinova D., Pavolv R., Lunchev D., Goynov M., Pavlova L.* Interactive game development to assist cultural heritage // Digital Presentation and Preservation of Cultural and Scientific Heritage. Vol. 8, Sofia, Bulgaria: Institute of Mathematics and Informatics – BAS. 2018. pp. 71-82.
25. *Paneva-Marinova D., Pavlov R.* Mini-symposium on future trends in serious games for cultural heritage // Digital Presentation and Preservation of Cultural and Scientific Heritage. vol. 8, Sofia, Bulgaria: Institute of Mathematics and Informatics – BAS. 2018. pp. 241-244.
26. *Zhonggen Y.* A meta-analysis of use of serious games in education over a decade // International Journal of Computer Games Technology. 2019.

УДК 004.65 **DOI** 10.15622/ia.20.4.7

В.В. Миронов, А.С. Гусаренко, Г.А. Тугузбаев

ИЗВЛЕЧЕНИЕ СЕМАНТИЧЕСКОЙ ИНФОРМАЦИИ ИЗ ГРАФИЧЕСКИХ СХЕМ

Миронов В.В., Гусаренко А.С., Тугузбаев Г.А. **Извлечение семантической информации из графических схем.**

Аннотация. Рассматривается задача извлечения семантической информации из электронного документа, заданного в формате векторной графики и содержащего графическую модель (схему), построенную с помощью графического редактора. Задача состоит в программном извлечении определенных структурных и параметрических свойств схемы и занесении их в базу данных для последующего использования. На основе проведенного анализа возможностей графических редакторов сделан вывод об актуальности этой задачи для универсальных редакторов, не привязанных к конкретным графическим нотациям и использующих открытые графические форматы документов, что допускает программную обработку. Предлагаемый подход рассматривает графические документы на трёх уровнях абстракции: концептуальном (семантические свойства схемы), логическом (представление семантических свойств на внутреннем уровне документа) и физическом (внутренняя организация графического документа). Решение задачи основано на построении концептуально-логического отображения, то есть отображения концептуальной модели схемы в логическую модель графического документа с учетом его физической модели. В рамках подхода разработан алгоритм построения указанного отображения, представленный в виде объектно-ориентированного псевдокода. Исследование внутренней разметки в открытых графических форматах позволило построить модели идентификации элементов схемы и их соединений между собой, что необходимо для конкретного применения алгоритма. Получены выражения для адресации элементов схемы и доступа к их свойствам. Предложенный подход реализован на основе ситуационно-ориентированной парадигмы, в рамках которой процесс извлечения управляется иерархической ситуационной моделью. Обрабатываемые данные задаются в ситуационной модели в виде виртуальных документов, отображаемых на разнородные внешние источники данных. Для решаемой задачи рассматривается отображение на два варианта форматов векторной графики: на «плоский» файл разметки и на набор таких файлов в электронном архиве. Практическое использование результатов иллюстрируется на примере извлечения семантической информации из графических моделей, разрабатываемых на различных этапах проектирования баз данных.

Ключевые слова: блок-схема, векторная графика, извлечения свойств, ситуационно-ориентированная парадигма, ситуационная модель, виртуальный документ.

1. Введение. В процессе разработки систем различного назначения широко применяются схемы (графические модели) различных видов (информационные, электрические, гидравлические, пневматические) и типов (структурные, функциональные, принципиальные и др.), которые с помощью условных графических обозначений показывают на том или ином уровне абстракции, какие компоненты содержит система и как они соотносятся между собой (ГОСТ 2.701–2008 ЕСКД. Схемы. Виды и типы. Общие требования к выполнению). В настоящее время схемы,

как правило, строятся разработчиками в виде электронных документов в формате векторной графики с помощью графических редакторов. В дальнейшем визуальное представление схемы используется как само по себе, так и в качестве основы для разработки других схем и иной конструкторской документации. В условиях стремления к автоматизации процессов разработки систем такие документы должны не только четко и ясно доносить до пользователей информацию в визуальной форме (презентационный аспект), но и предоставлять формальные спецификации системы, заложенные в схеме (семантический аспект).

В связи с этим возникает задача извлечения формальных спецификаций (семантической информации) из графического документа, которая рассматривается в данной статье. Ниже анализируется состояние области исследования, обсуждается задача извлечения и предлагается подход к ее решению. Предлагается модель, задающая отображение концептуальной модели схемы в логическую модель графического документа, и рассматривается общий алгоритм задания концептуальной модели. Обсуждается идентификация элементов схемы для различных графических форматов. Описывается реализация предложенного подхода на основе ситуационно-ориентированной парадигмы и применение результатов в учебном процессе.

2. Состояние области исследования. Графические редакторы, применяемые для построения графических моделей, можно разделить на специализированные и универсальные.

Специализированные редакторы ориентированы на создание схем (диаграмм) в узкой предметной области. Например, популярный редактор Erwin Data Modeler дает возможность построения ER-диаграмм баз данных в нотации IDEF1X, редактор IBM Rational Rose XDE предоставляет разработчику среду разработки программного обеспечения в нотации UML. Эти редакторы имеют специфическую функциональность, такую как генерация диаграмм из программного кода в сочетании с генерацией программного кода из диаграмм (roundtrip engineering). В доступных научных публикациях исследуется повышение надежности автоматизированных систем в ходе трансформации диаграмм за счет сегментирования модели [1]. Обсуждаются программные средства для автоматического создания визуальных композиций на основе диаграмм [2]. Рассматриваются инструментарии, допускающие использование информационных сообщений, встроенных в графику [3, 4]. Анализируются преимущества и недостатки наиболее популярных редакторов объектно-ориентированного моделирования [5].

В целом специализированные редакторы ориентированы на жестко заданные графические нотации (системы условных обозначений и правил записи схем) и предоставляют ограниченный набор сведений о свойствах схемы (таких, как статистика — количество элементов, связей и т. п.). С их помощью нельзя задавать схемы в непредусмотренной нотации и нельзя извлекать семантические свойства схемы, не предусмотренные исходной функциональностью.

Универсальные графические редакторы не ограничены какой-либо графической нотацией. Наборы фигур, позволяющие задавать условные графические обозначения в той или иной нотации, поставляются вместе с редактором или сторонними разработчиками, а также могут быть созданы самими пользователями. Популярными универсальными редакторами векторной графики являются Microsoft Office Visio (проприетарный), а также его свободно распространяемые аналоги OpenOffice Draw, LibreOffice Draw и др. [6]. В многочисленных публикациях отражается эффективное применение этих редакторов для создания схем в самых разных предметных областях: учебном процессе [7, 8], разработке и документировании программного обеспечения [9], моделировании бизнес-процессов [10], проектировании автоматизированных систем [11], разработке и анализе моделей рабочих процессов [12–14], интеграции баз данных [15–17], обработке запросов на естественном языке [18] и др.

Универсальные редакторы не привязаны к конкретным графическим нотациям, поэтому с их помощью можно задавать схемы в заранее не предусмотренной нотации. Другой их особенностью является использование открытых графических форматов для создания результирующих схем. Это дает принципиальную возможность автоматизированной обработки графических документов с целью извлечения семантических свойств из содержащихся в них схем. Вместе с тем, как показал анализ публикаций, вопросы извлечения свойств схемы практически не освещены в литературе (ни в техническом, ни в научном аспектах). Отсутствуют как общие принципы, так и практические аналоги решения этой задачи. Это обуславливает актуальность исследования по извлечению семантических свойств схемы применительно к универсальным графическим редакторам.

3. Задача извлечения. Целью исследования являлось достижение понимания того, как может быть извлечена семантическая информация из схемы, содержащейся в электронном документе, созданном в среде универсального графического редактора в формате векторной графики. Под семантической информацией здесь понимаются структурные и па-

раметрические свойства схемы, очищенные (абстрагированные) от несущественных деталей ее визуального представления, таких как размеры фигур, толщина линий, цвет заливки и т. п. Результатом решения задачи должен быть граф, вершины которого соответствуют элементам схемы, а дуги — соединениям элементов друг с другом. Вершины и дуги должны быть нагружены параметрами, заданными в схеме. В техническом аспекте задача состоит в программном извлечении интересующих структурных и параметрических свойств графической схемы и занесение их в базу данных для последующего использования. При этом следует различать два уровня представления информации в электронном графическом документе: внешний, соответствующий визуальному отображению схемы, и внутренний, соответствующий внутреннему кодированию элементов схемы согласно используемому графическому формату.

Идея решения задачи состоит в том, чтобы исследовать особенности задания условных обозначений на внутреннем уровне в различных форматах векторной графики и на этой основе научиться идентифицировать условные обозначения, примененные в конкретной схеме. Предлагаемый подход в общем виде иллюстрируется на рисунке 1.

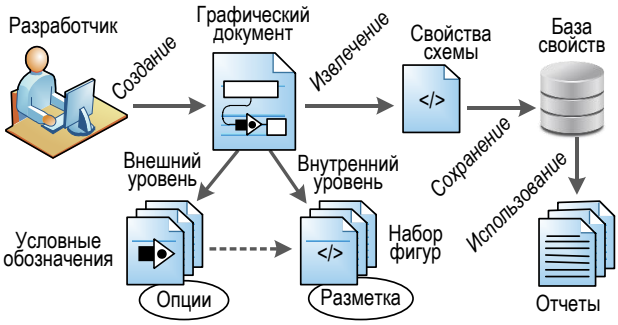


Рис. 1. Подход к извлечению семантической информации из электронного графического документа

Создание. Предполагается, что электронный документ содержит блок-схему, построенную в определенной графической нотации на основе определенного набора фигур-элементов (условных обозначений). Разработчику схемы в среде графического редактора (как специализированного, так и универсального) предоставляется набор фигур, задающих условные обозначения. Условным обозначениям на внешнем, визуальном уровне соответствует набор закодированных фигур изображения

на внутреннем уровне документа. При этом установленные разработчиком опции (допустимые модификации) условных обозначений определенным образом отражаются во внутреннем представлении (разметке) набора фигур. Использование стандартного набора фигур на внешнем, визуальном уровне дает возможность идентифицировать как сами фигуры, так и их опции при последующей программной обработке документа на внутреннем уровне. Поэтому возникает важный вопрос о том, каким образом условные обозначения, используемые для визуального представления схемы, реализуются на внутреннем уровне графического документа.

Извлечение. Возможность программной обработки графического документа с целью извлечения из него свойств обусловлена применением в универсальных графических редакторах открытых форматов. Открытые графические форматы, как правило, основаны на использовании XML-разметки и двух способов организации:

• плоский формат, при котором графический документ представляет собой несжатый XML-файл. Например, формат VDX (Visio Drawing based on XML) представляет XML-файл графического редактора Visio на языке разметки DataDiagramML. Этот формат применяется в версии Microsoft Visio 2010, он достаточно распространен, хотя в поздних версиях Visio не поддерживается. Или формат FODG (Flat ODG), применяемый в редакторе LibreOffice Draw — свободно распространяемом аналоге редактора Visio;

• иерархический формат, продвинутый, при котором графический документ организован в виде сжатой в ZIP-архиве иерархии вложенных папок, содержащих XML-файлы компонентов документа. Например, формат VSDX в редакторе Visio, соответствующий стандарту Open Packaging Conventions и применяемый в версиях Visio 2013 и более поздних. Или формат ODG, соответствующий стандарту Open Document Format и применяемый в свободно распространяемых редакторах Apache OpenOffice Draw и LibreOffice Draw.

Сохранение. Результаты извлечения необходимо сохранить в базе данных в виде, удобном для дальнейшего использования. Извлечение предполагает абстрагирование, то есть получение из графического документа информации о существенных структурно-параметрических свойствах схемы, игнорируя при этом несущественные стилистические и эстетические особенности графического изображения. К структурно-параметрическим свойствам схемы, которые нужно выявить в результате извлечения, относятся:

• элементы — экземпляры условных графических обозначений, использованные в схеме;

- установленные опции и параметры каждого элемента схемы;
- соединение элементов между собой с помощью коннекторов.

Использование. Результаты извлечения свойств могут быть использованы для различных целей, таких как проверка корректности схемы, выявление ошибок; генерация проектной документации; генерация программного кода, соответствующего схеме; формирование персонализированных моделей для следующих этапов проектирования.

Отметим, что рассматриваемая задача извлечения ориентирована не на какую-то конкретную схему, а на множество схем определенного типа. То есть графическая схема здесь рассматривается как тип или класс, предполагающий множество конкретных экземпляров его реализации. Все экземпляры выполнены в одной графической нотации, то есть основываются на общей системе условных обозначений. При этом при построении экземпляров схемы применяется один и тот же набор фигур, реализующих условные обозначения. Экземпляры схемы отличаются только составом фигур, их опциями и параметрами, предусмотренными графической нотацией, а также соединениями фигур между собой с помощью коннекторов. Такие допущения позволяют единообразно обрабатывать экземпляры схемы в процессе извлечения свойств.

Отметим также, что задача извлечения может относиться не ко всей схеме, а к некоторой ее части. Она может касаться, к примеру, только определенных свойств или аспектов схемы и/или относиться только к определенным частям или элементам схемы.

4. Модель концептуально-логического отображения. В основе предлагаемого подхода используется рассмотрение графического документа на трёх уровнях абстракции: концептуальном (семантические свойства схемы), логическом (представление семантических свойств на внутреннем уровне документа) и физическом (внутренняя организация графического документа). Решение задачи основано на построении концептуально-логического отображения, то есть построении модели, задающей отображение концептуальной модели схемы в логическую модель графического документа с учетом его физической модели (рис. 2).

Концептуальная модель. Под концептуальной моделью схемы нами понимается модель верхнего уровня абстракции, задающая набор элементов, их свойства, а также отношения друг с другом. Каждый элемент схемы — это экземпляр условного графического изображения. Концептуальная модель задает, какие элементы присутствуют на схеме; каковы свойства (опции, параметры) этих экземпляров; как экземпляры вложены друг в друга; как экземпляры соединены между собой. На ри-

сунке 2 справа концептуальная модель иллюстрируется с помощью дерева, где корень Schema содержит множество дочерних элементов (Element) и множество дочерних соединений (Connect), которые, в свою очередь, содержат множества свойств (Property). Отношение вложенности одного элемента в другой задается необязательной ссылкой на родителя (Parent). Отношение соединения одного элемента с другим задается парой ссылок на эти элементы (From и To), которые содержатся в соединении. Таким образом, концептуальная модель отражает структуру того, что должно быть получено в результате извлечения.

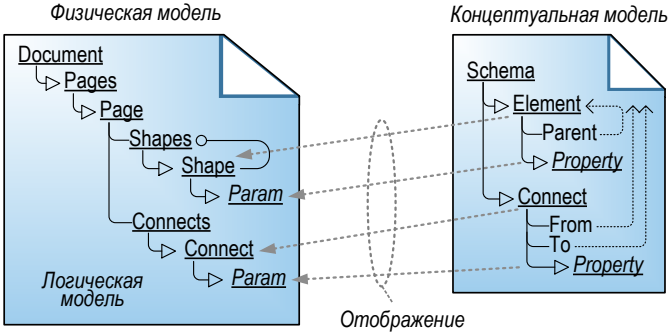


Рис. 2. Концептуально-логическое отображение

Физическая модель. Физическая модель в нашем понимании — это общая модель графических документов с учетом используемого графического формата и используемого набора фигур условных обозначений. Одному и тому же внешнему (визуальному) представлению схемы соответствуют разные физические модели при использовании разных графических форматов, разных наборов фигур для представления элементов (условных обозначений), а также разного визуального представления и размещения фигур на листах документа. Экземплярами этой модели являются внутренние представления (в виде XML-деревьев) конкретных графических документов, построенных с использованием определенного набора фигур, представляющих элементы определенной графической нотации. Таким образом, физическая модель соответствует исходному документу, содержащему схему, которую нужно подвергнуть извлечению.

Логическая модель. Под логической моделью здесь понимается та часть физической модели, которая имеет отношение к элементам

и свойствам схемы, отраженным в концептуальной модели. В логической модели опущены свойства, задающие внешнее, визуальное представление схемы (такие как размеры элементов, толщина линий, координаты положения на листе, цвет, заливка и др., если только они не существенны с точки зрения концептуальной модели). Как правило, в логической модели игнорируется значительное количество свойств внешнего визуального представления документа, поэтому логические модели существенно проще физических. Однако и они зависят от графического формата и способа задания условных обозначений. На рисунке 2 слева логическая модель иллюстрируется с помощью дерева, являющегося XML-поддеревом физической модели. Корневой XML-элемент Document содержит дочерний XML-элемент Pages, который, в свою очередь, включает множество дочерних XML-элементов Page, задающих изображение на отдельных листах графического документа (префикс XML введен, чтобы отличать понятие «элемента» в языке XML от элемента схемы). Вложенный XML-элемент Shapes включает множество дочерних XML-элементов Shape, задающих отдельные фигуры изображения. Фигуры могут быть сложными, состоящими из других фигур. В этом случае применяется рекурсивная вложенность: внутри XML-элемента Shape размещается XML-элемент Shapes с вложенными Shape, задающими детализацию. Соединения фигур на листе задаются с помощью XML-элемента Connects, включающего множество дочерних XML-элементов Connect, каждый из которых определяет пару соединенных фигур Shape, которые ранее уже определены. Таким образом, логическая модель отражает те особенности исходного графического документа, которые существенны в задаче извлечения.

Концептуально-логическое отображение. Для решения задачи необходимо построить концептуально-логическое отображение, которое ставит в соответствие объектам концептуальной модели (элементам, соединениям и их свойствам) объекты логической модели (а также физической модели, поскольку логическая модель есть часть физической). Тем самым концептуально-логическое отображение дает понимание того, где в графическом документе находится XML-разметка фигур, составляющих тот или иной элемент схемы, и как параметры этой разметки задают свойства (опции и параметры) элемента схемы.

На рисунке 3 представлена модель «сущность–связь», задающая основные типы сущностей, участвующих в задаче извлечения, (фигуры-прямоугольники) и бинарные связи между ними (фигуры-стрелки).

Связи указывают направление от сущности-родителя к сущности-ребенку с кардинальностью «0, 1, M» (треугольник — связь «ко

многим») или «0, 1» (трапеция — условная связь). Кружок внутри треугольника задает кардинальность связи в направлении от ребенка к родителю: темный — «1, 1» (недопустимы «сироты» — каждый ребенок должен иметь родителя); светлый — «0,1» (допустимы «сироты» — ребенок может не иметь родителя). Темный квадратик обозначает идентифицирующую связь (когда идентификатор родителя входит в состав идентификатора ребенка), а светлый — неидентифицирующую (когда ребенок идентифицируется независимо от родителя).

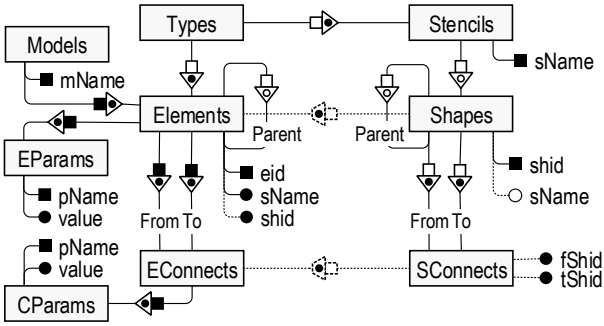


Рис. 3. Взаимосвязь сущностей на концептуальном (слева) и логическом/физическом (справа) уровнях

Графическая нотация. Сущность типа `Types` задает множество типов элементов, то есть типов условных графических изображений графической нотации, применяемой в схеме. Каждый экземпляр этой сущности соответствует конкретному типу. Множество экземпляров `Types` для задачи извлечения известно априори.

Сущность типа `Stencils` задает множество трафаретов (шаблонов, образцов) фигур, соответствующих условным графическим обозначениям, которые используются разработчиком для построения схемы путем копирования на лист графического документа в среде графического редактора. Наборы трафаретов могут существовать в различных видах: в виде фигур, хранящихся в отдельном графическом документе, или (если позволяет функциональность графического редактора) в виде специальных наборов фигур, которые могут прикрепляться к графическому документу и использоваться для его построения. Каждому экземпляру сущности `Stencils` на внутреннем уровне соответствует некоторое дерево XML-разметки. Экземпляр трафарета идентифицируется ключевым атрибутом `sName` («имя трафарета») и относится к определенному типу, то есть ему соответствует один и только один экземпляр

сущности Types. Вместе с тем возможно несколько трафаретов одного типа, например, которые отличаются визуальным стилем исполнения или конкретными значениями параметров соответствующего условного графического обозначения.

Графический документ. Сущность типа Shapes задает множество фигур физической модели схемы. Каждый экземпляр этой сущности соответствует конкретной фигуре, присутствующей в графическом документе. На внутреннем уровне экземпляр соответствует поддереву XML-разметки XML-дерева документа. Экземпляр фигуры идентифицируется ключевым атрибутом shid («идентификатор фигуры»). Каждому экземпляру фигуры, если он получен путем копирования некоторого трафарета, соответствует один экземпляр сущности Stencils, этот факт отражен с помощью виртуального атрибута sName, значение которого копируется из сущности Stencils. Вместе с тем в графическом документе возможны фигуры, не относящиеся к какому-либо трафарету. Рекурсивная связь Parent задает вложенность фигур — каждый экземпляр сущности Shapes может иметь (или не иметь) родителя — экземпляр сущности Shapes, в который он вложен; при этом каждый экземпляр сущности Shapes может иметь ноль или несколько вложенных дочерних экземпляров Shapes.

Сущность типа SConnects задает множество пар соединенных друг с другом фигур, то есть каждому экземпляру этой сущности соответствует пара экземпляров сущности Shapes: один экземпляр, являющийся первым (From), и один экземпляр, являющийся вторым (To) компонентом пары. Из экземпляра соединения доступны, таким образом, идентификаторы двух родителей, представленные на рисунке в виде виртуальных атрибутов fShid и tShid. При этом каждая фигура может участвовать в ноль или нескольких экземплярах соединений в качестве как первого, так и второго компонента.

Результирующая концептуальная модель. Сущность типа Models задает множество концептуальных моделей схемы. В общем случае из одного графического документа можно получить несколько концептуальных моделей, отличающихся детальностью или полнотой извлекаемых сведений. Экземпляры этой сущности идентифицируются ключевым атрибутом mName («имя модели»).

Сущность типа Elements задает множество элементов концептуальной модели схемы. Каждый экземпляр этой сущности соответствует конкретному элементу, присутствующему на схеме, и принадлежит определенной модели (экземпляру сущности Models). Экземпляры этой

сущности идентифицируются ключевым атрибутом `eid` («идентификатор элемента»). Атрибут `sName` задает имя трафарета, использованного при задании соответствующей фигуры в графическом документе, а атрибут `shid` задает идентификатор самой фигуры. Элементы схемы относятся к определенному типу, то есть каждому экземпляру сущности `Elements` соответствует один и только один экземпляр сущности `Types`. Рекурсивная связь `Parent` задает вложенность элементов — каждый экземпляр сущности `Elements` может иметь (или не иметь) родителя — экземпляр сущности `Elements`, в который он вложен; с другой стороны, каждый экземпляр сущности `Elements` может иметь ноль или несколько вложенных дочерних экземпляров того же типа. Множество экземпляров сущности `Elements` должно быть построено в результате решения задачи извлечения.

Сущность типа `EParams` задает множество параметров элементов схемы. Каждый экземпляр этой сущности соответствует конкретному параметру конкретного элемента, присутствующего на схеме. Экземпляры этой сущности в пределах одного родительского элемента идентифицируются ключевым атрибутом `pName` («имя параметра»). Атрибут `value` задает значение параметра. Множество экземпляров `EParams` должно быть построено в результате решения задачи извлечения.

Сущность типа `EConnects` задает множество пар соединенных друг с другом элементов схемы, то есть каждому экземпляру этой сущности соответствует пара экземпляров сущности `Elements`: один экземпляр, являющийся первым (`From`), и один экземпляр, являющийся вторым (`To`) компонентом пары. Экземпляры этой сущности идентифицируются парой ключевых атрибутов родителей через связи `From` и `To`. При этом каждый элемент может участвовать в ноль или нескольких экземплярах соединений в качестве как первого, так и второго компонента. Каждому экземпляру сущности `EConnects` соответствует экземпляр сущности `SConnects`, но не наоборот. Множество экземпляров сущности `EConnects`, как и множество экземпляров сущности `Elements`, должно быть построено в результате решения задачи извлечения.

Сущность типа `CParams` задает множество параметров соединений схемы. Каждый экземпляр этой сущности соответствует конкретному параметру конкретного соединения из множества `EConnects`. Экземпляры этой сущности в пределах одного родительского соединения идентифицируются ключевым атрибутом `pName` («имя параметра»). Атрибут `value` задает значение параметра. Множество экземпляров сущности `CParams` должно быть построено в результате решения задачи извлечения.

Допущения. Отметим существенное допущение, принятое в этой модели, а именно, предположение, что каждому элементу схемы соответствует некоторая (единственная) фигура внутреннего представления. На рисунке 3 это обстоятельство отмечено темным кружком на символе условной связи `Shapes-Elements` (эта связь изображена пунктиром, поскольку она только подразумевается, ее реализация не требуется явно). Условные графические изображения сами по себе могут быть сложными, состоящими из множества графических примитивов, их приходится рисовать средствами графического редактора на основе сочетания нескольких фигур. Указанное допущение требует, чтобы в этом случае все фигуры-компоненты, составляющие элемент схемы как условное графическое обозначение, были заключены в единую фигуру-контейнер, представляющую этот элемент. Разработчик схемы должен не просто нарисовать элемент схемы в соответствии с визуальными правилами выполнения условных графических изображений, но оформить его в виде единой внутренней фигуры. Графические редакторы предоставляют для этого возможность группирования нескольких фигур изображения в единую фигуру-контейнер.

Другое допущение связано с соединением фигур. Разработчик должен не просто зрительно разместить рядом соединенные элементы, но закрепить это соединение так, чтобы этот факт отразился особым образом во внутреннем представлении графического документа, мог быть обнаружен в ходе программной обработки и зафиксирован в концептуальной модели в виде соответствующего экземпляра сущности `EConnects`. Графические редакторы предоставляют для этого возможность размещения в фигурах точек соединения, линий-коннекторов, а также предусматривают специальный механизм фиксации соединений.

5. Общий алгоритм задания концептуальной модели. Итак, множества экземпляров сущностей `Types` и `Stencils` заданы априори для используемой графической нотации. Множество экземпляров сущностей `Shapes` и `SConnects` задано в XML-разметке обрабатываемого графического документа. Множество экземпляров сущностей `Elements` и `EConnects` требуется определить в результате извлечения на основе анализа экземпляров сущностей `Shapes` и `SConnects`. Для этого должна быть возможность идентифицировать фигуры графического документа на предмет их соответствия элементам схемы определенного типа, а также идентифицировать параметры элементов, основываясь на свойствах XML-разметки соответствующих фигур. В соответствии с этим был разработан алгоритм задания результирующей концептуальной модели схемы, объектно-ориентированный псевдокод представлен ниже в листинге 1.

```
000 Begin
001   m = New Model ('MName')
002   ForEach s: s ∈ Shapes, s.sName ∈ StencilNames
003      e = m.AddElement (s.shid)
004      e.AddType (EType (s.sName))
005      ForEach ep: ep ∈ EParamNames (s.cName)
006         e.AddEParam (p, s.pvalue (p))
007   ForEach c: c ∈ SConnects
008      For fe: fe ∈ m.Elements, fe.shid = c.fShid
009         For te: te ∈ m.Elements, te.shid = c.tShid
010            cm = m.AddEConnect (fe, te)
011            ForEach cp: cp ∈ CParamNames (fe.sName, te.sName)
012               cm.AddCParam (cp, c.pvalue (cp))
013 End
```

Листинг 1. Псевдокод алгоритма задания концептуальной модели

Строки `000` и `013` задают контейнер для кода в целом. В строке `001` создается новая (пока еще пустая) результирующая модель в виде объекта `m`; имя модели задается с помощью параметра `'MName'`.

Фрагмент `002`–`006` обрабатывает фигуры графического документа и заполняет модель обнаруженными элементами и их параметрами. Конструкция `ForEach` (строка `002`) перебирает все фигуры графического документа, отбирая те из них, которые основаны на каком-либо трафарете используемой графической нотации. В качестве критерия проверяется принадлежность атрибута `sName` фигуры множеству имен используемых трафаретов `StencilNames`. Далее для каждой отобранной фигуры `s` выполняется следующее:

- (строка `003`) в модель `m` добавляется новый элемент `e`, соответствующий фигуре `s`, идентификатор фигуры `shid` передается в качестве параметра;
- (строка `004`) для элемента `e` устанавливается ссылка на его тип, для чего используется функция `EType`, отображающая имена шаблонов в типы элементов;
- (строки `005`–`006`) формируется список параметров элемента `e`, для чего перебираются имена параметров `EParamNames`, предусмотренных для трафарета `s.sName` элемента `e`, и с помощью метода `AddEParam`

в объект e добавляются новые объекты-параметры с заданными именами pName и значениями value. Значение параметра формируется с помощью функции pvalue на основе XML-разметки фигуры s;

Фрагмент 007–012 обрабатывает соединения фигур графического документа и заполняет модель соединениями элементов и параметрами соединений. Конструкция ForEach (строка 007) перебирает все соединения фигур графического документа. Следующие конструкции For (строки 008 и 009) отбирает те соединения, которые относятся к фигурам-элементам и отыскивает соответствующую пару элементов fe (первый) и te (второй). Далее для каждой отобранной пары элементов выполняется следующее:

• (строка 010) в модель m добавляется новое соединение элементов cm;

• (строки 011–012) формируется список параметров соединения элементов fe и te, для чего перебираются имена параметров CParamNames, предусмотренных для соответствующей пары трафаретов, и с помощью метода AddCParam в объект cm добавляются новые объекты-параметры с заданными именами pName и значениями value. Значение параметра формируется с помощью функции pvalue на основе XML-разметки соединения cp.

6. Идентификация элементов схемы. При обработке графического документа согласно рассмотренному алгоритму, необходимо выполнить перебор составляющих его фигур с целью идентификации элементов схемы. Этот процесс зависит, во-первых, от используемого графического формата, во-вторых, от конкретного представления фигур условных графических обозначений. Таким образом, возникает задача построения моделей идентификации элементов схемы для конкретных графических форматов. Рассматриваемые ниже модели получены в результате непосредственного исследования внутренней XML-разметки графических документов для открытых форматов редакторов Visio, LibreOffice Draw и OpenOffice Draw.

Особенности идентификации трафарета фигуры. Для каждой обрабатываемой фигуры s необходимо узнать значение s.sName, то есть соответствует ли фигура одному из трафаретов и каково уникальное имя sName этого трафарета. Рассмотрим организацию именования трафаретов и фигур в различных графических форматах.

Форматы VDX/VSDX (графический редактор Visio). В этих форматах сложные фигуры существуют в документе в двух формах: 1) в форме так называемого мастера (Master — «хозяин») — исходного

трафарета фигуры; 2) в форме фигуры-экземпляра (`Shape`), представляющего конкретный экземпляр мастера-трафарета, который размещен на листе со своими индивидуальными настройками. Один мастер-трафарет может служить основой для множества фигур-экземпляров. Имя мастера задается разработчиком при визуальном создании фигуры и сохраняется во внутреннем XML-представлении мастера (атрибут `NameU` — «универсальное имя» — у XML-элемента `Master`). Внутренняя XML-разметка фигуры-экземпляра не всегда содержит атрибут `NameU`, но всегда содержит ссылку на свой мастер-трафарет, если фигура была порождена на основе трафарета. Поэтому имя трафарета фигуры может быть идентифицировано по имени `NameU` его мастера. Реализация этого подхода различается для форматов VDX и VSDX.

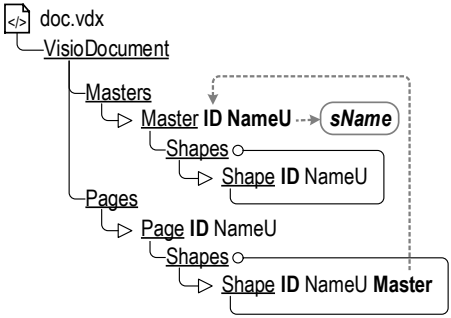


Рис. 4. Извлечение имени трафарета фигуры в формате VDX

В первом случае мастера размещаются в том же XML-файле, что и порожденные из них фигуры (рис. 4). Мастера размещаются в элементах `/VisioDocument/Masters/Master` и идентифицируются уникальным XML-атрибутом `ID`. Фигуры размещаются в элементах `/VisioDocument/Pages/Page/Shapes/Shape` и имеют XML-атрибут `Master`, содержащий значение идентификатора мастера. Таким образом, чтобы получить имя трафарета, нужно извлечь значение атрибута `NameU` у мастера, идентификатор которого совпадает со значением атрибута `Master` обрабатываемой фигуры. На языке XPath (языке адресации узлов в XML-деревьях) это будет выглядеть, например, так:

```
//Master[@ID = ./@Master]/@NameU.
```

При этом предполагается, что контекстным узлом, обозначенным как «.», является обрабатываемая фигура `Shape`.

Во втором случае список мастеров, их описания, а также содержимое отдельных листов документа размещено в отдельных XML-файлах в составе ZIP-архива (рис. 5).

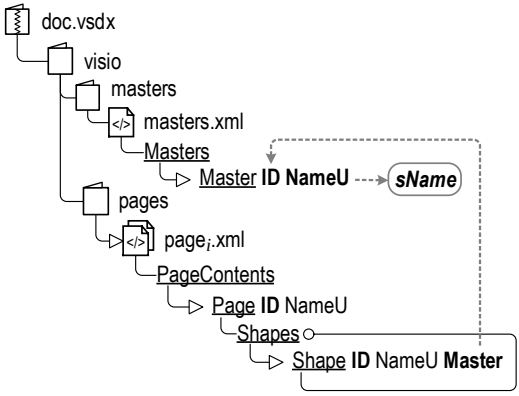


Рис. 5. Получение имени трафарета фигуры в формате VSDX

Здесь общие сведения о мастере, в том числе и необходимый нам атрибут `NameU`, находятся в элементах `/Masters/Master` файла `masters.xml`, который размещен в папке `masters` ZIP-архива графического документа. Сведения о фигуре находятся в элементах `/PageContents/Page/Shapes/Shape` файла `page`*i*`.xml`, который соответствует *i*-му листу графического документа (размещен в папке `Pages`). Таким образом, извлечение имени трафарета в этом случает требует совместной обработки двух XML-файлов.

Формат ODG (графические редакторы LibreOffice / OpenOffice Draw). В этом формате, как и во многих других, используемых в графических редакторах — аналогах Visio, внутреннее представление графического документа не предусматривает отдельного сохранения трафарета-мастера — исходного представления скопированного изображения. XML-разметка фигуры просто копируется в список фигур страницы документа. Однако в составе внутреннего XML-представления скопированных фигур присутствует необязательный XML-элемент `svg:title` (рис. 6), который может быть задан разработчиком при создании исходного изображения фигуры-трафарета и который наследуется во всех его копиях. Этот XML-элемент может быть использован в качестве имени трафарета.

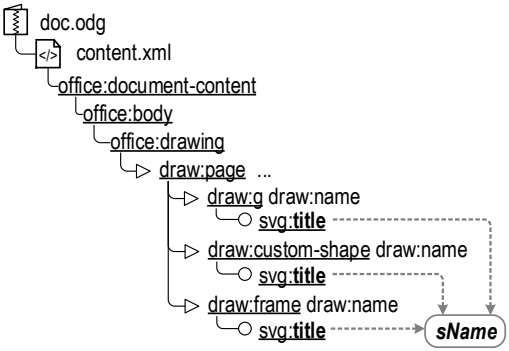


Рис. 6. Получение имени трафарета фигуры в формате ODG

В разметке листа `draw:page` графического документа формата ODG конкретные фигуры могут быть представлены различными XML-элементами: `draw:g`, `draw:custom-shape`, `draw:frame` (рис. 6). Фигуры идентифицируются уникальными именами `draw:name`, а вложенные XML-элементы `svg:title` содержат искомое имя трафарета. Таким образом, в этом случае имя трафарета фигуры содержится в XML-разметке самой фигуры, что облегчает идентификацию. Соответствующее XPath-выражение выглядит так:

```
./svg:title .
```

Для сравнения на рисунке 7 приведена аналогичная модель для документа в формате GraphML, который применяется, в частности, в популярном редакторе yEd Graph Editor.

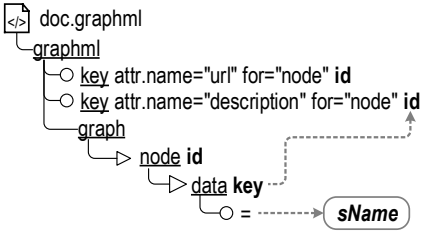


Рис. 7. Получение имени трафарета фигуры в формате GraphML

Здесь отдельные фигуры заданы с помощью XML-элементов `/graphml/graph/node`, а их содержимое задается с помощью вложенных XML-элементов `data`. Каждый элемент `data` ссылается на XML-элемент

key, поясняющий его назначение. В частности, элементы key с атрибутами attr.name = "url" или attr.name = "description", соответствуют элементам data, содержащим пользовательские данные. Их можно использовать для задания имени фигуры-трафарета. Так, если имя трафарета задано разработчиком в поле description, то его можно извлечь как текстовое содержимое соответствующего элемента data, например, с помощью следующего XPath-выражения:

```
./data[@key=//key[@attr.name="description"]/@id] .
```

Таким образом, идентификация трафарета фигуры существенно зависит от используемого графического формата.

Особенности идентификации соединений. Идентификация соединений требует выяснить на основе анализа соединений фигур графического документа, какие элементы схемы, выявленные и внесенные в концептуальную модель ранее, соединены между собой. Это несложно сделать при совместной обработке графического документа и концептуальной модели, поскольку внесенные элементы схемы содержат ссылки на соответствующие им фигуры графического документа (рис. 4, листинг 1). Однако детали зависят от используемого графического формата (рис. 8).

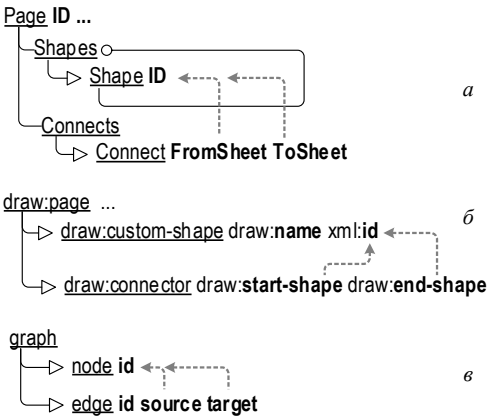


Рис. 8. Организация соединения фигур в различных форматах: а - VDX/VSDX, б - ODG, в - GraphML

Форматы VDX/VSDX. В этих форматах каждое соединение пары фигур отражается в XML-элементе Connects/Connect, который вложен в XML-элемент Page. Атрибуты FromSheet и ToSheet XML-элемента

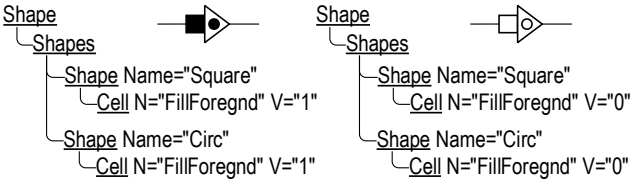
`Connect` ссылаются на идентификаторы первой и второй соединяемых фигур `Shapes/Shape` (рис. 8, *a*).

Форматы ODG и GraphML. В этих форматах каждое соединение отражается с помощью соответствующих атрибутов в фигурах-коннекторах. В формате ODG каждый коннектор задается XML-элементом `draw:connector`, вложенным в XML-элемент `draw:page`. Атрибуты `draw:start-shape` и `draw:end-shape` XML-элемента `draw:connector` ссылаются на идентификаторы первой и второй фигур, соединяемых данным коннектором (рис. 8, *b*). В формате GraphML каждый коннектор задается XML-элементом `edge`, вложенным в XML-элемент `graph`. Атрибуты `source` и `target` XML-элемента `edge` ссылаются на идентификаторы первой и второй фигур, соединяемых коннектором (рис. 8, *c*).

Идентификация параметров элемента схемы. Идентификация параметров элемента должна ответить на вопрос о том, какие особенности имеет данный элемент схемы как конкретный экземпляр условного графического обозначения определенного типа. Идентификация параметров выполняется путем исследования внутреннего представления фигуры, соответствующей элементу. Поскольку состав параметров определяется используемым трафаретом элемента, идентификация параметров зависит от результата предшествующей идентификации имени трафарета элемента, то есть для каждого трафарета элемента используется свой алгоритм идентификации того или иного параметра. Значение определенного параметра элемента определяется особенностями внутренней XML-разметки фигуры этого элемента — текстовыми значениями определенных XML-атрибутов или вложенных XML-элементов или их сочетаний. Для определения значения параметра строится XPath-выражение, которое адресует соответствующие XML-атрибуты или элементы, позволяя извлекать их значения. Аналогичным образом идентификация параметров соединения выполняется путем исследования внутреннего представления соединения, а также внутреннего представления соединяемых фигур. Для концептуальной модели схемы может быть существенным то, к каким частям фигур или к каким точкам соединения прикреплены коннекторы.

Пример. В качестве примера на рисунке 9 приведены модели XML-разметки двух вариантов задания элемента «связь», используемого при задании моделей базы данных на концептуальном уровне абстракции (модель «сущность–связь») в так называемой «уфимской» нотации. Символ связи представляет собой сочетание фигур-компонентов: квадрата, треугольника, круга и двух коннекторов. Параметры связи отображаются цветом заливки фигур-компонентов. Например, темная заливка квадрата означает идентифицирующую связь, а светлая

— не идентифицирующую; темная заливка круга означает обязательную связь от ребенка к родителю, а светлая — необязательную. Внутренняя разметка `Shape` фигуры «связь» в формате VSDX содержит разметку отдельных фигур-компонентов `Shape`, вложенных в контейнер `Shapes`.



а) *б)*

Рис. 9. Пример задания параметров условного графического обозначения «связь» в формате VSDX: а) идентифицирующая обязательная; б) неидентифицирующая необязательная

На рисунке 9 для примера представлена разметка двух фигур-компонентов: квадрата с атрибутом `Name="Square"` и круга с атрибутом `Name="Circ"`. Заливка фигур в формате VSDX задается с помощью XML-элементов `Cell` с атрибутом `N="FillForegnd"`, при этом значение цвета заливки задается атрибутом `V`. Таким образом, значение цвета заливки квадрата можно получить с помощью следующего XPath-выражения:

```
./Shapes/Shape[@Name="Square"]/Cell[@N="FillForegnd"]/@V ,
```

где предполагается, что контекстным узлом, обозначенным как «.», является обрабатываемая фигура «связь» `Shape`. Соответственно, значение цвета заливки круга можно получить с помощью XPath-выражения

```
./Shapes/Shape[@Name="Circ"]/Cell[@N="FillForegnd"]/@V .
```

7. Реализация на основе ситуационно-ориентированной парадигмы. Исследование процесса извлечения, рассмотренное выше, имело в качестве побочной цели развитие ситуационно-ориентированного подхода в плане распространения его на обработку документов векторной графики. Ситуационно-ориентированный подход исследуется авторами в рамках проекта интеграции разнородных данных на основе иерархической ситуационной модели (HSM — Hierarchical Situation Model) [18]. HSM определяет функционирование ситуационно-ориентированной базы данных на основе мониторинга текущей ситуации и сохранения текущего состояния ситуации между сеансами интерпретации. Архитектура HSM-модели иллюстрируется на рисунке 10.

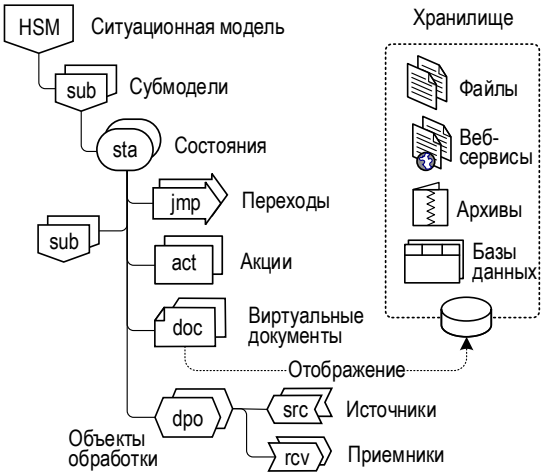


Рис. 10. Архитектура HSM

Субмодели `sub`, составляющие модель HSM, образуют иерархию — каждое состояние `sta` в субмодели в свою очередь может содержать вложенные субмодели. Элементы `jmp` задают переходы состояний, а элементы `act` — действия, ассоциированные с состояниями. Для рассматриваемой задачи существенными являются разновидности акций `doc` (виртуальные документы) и `dpo` (объекты обработки). В процессе сеанса интерпретации HSM контролируются текущие состояния субмоделей и выполняются ассоциированные акции, в том числе задающие обработку внешних данных. Для доступа к внешним данным применяются виртуальные документы `doc`, которые могут отображаться на внешние данные гетерогенной природы, исследованные в работах [18–20] (рис. 10): локальные и удаленные файлы, веб-сервисы, архивы, реляционные базы данных. Для обработки данных в объектах `dpo` предусмотрены элементы `src` (источники), задающие загрузку внешних данных, и элементы `rcv` (приемники), обеспечивающие вывод результатов. Эти элементы включают спецификации обработки данных.

В рассматриваемой задаче для нас важна возможность задавать в HSM отображение виртуальных документов на графические документы в различных открытых графических форматах. В работе [21] авторы рассматривали подобное отображение с целью генерации персо-

нализированных заготовок графических документов; в этой статье задача рассматривается в аспекте извлечения концептуальных свойств графического изображения.

Возможность отображения виртуальных документов на открытые графические форматы обусловлена тем, что все они в основе используют XML-разметку и организованы либо в виде плоского XML-файла, либо в виде архива XML-файлов. В первом случае следует использовать возможность отображения виртуального документа на «плоский» XML-файл, а во втором — на ZIP-архив XML-файлов. Например, HSM-декларация

```
<doc:VdxDocument type = "xml" path = "vdx/doc123.vdx/>"
```

задает отображение виртуального документа `doc:VdxDocument` целиком на VDX-документ `doc123.vdx`, находящийся в папке `vdx`. Декларация

```
<doc:VsdxDocument type = "zip" path = "vdx/doc456.vsdx">
    <ent:Master1 path = "visio/masters/master1.xml"/>
    <ent:Page1 path = "visio/pages/page1.xml"/>
</doc:VsdxDocument>
```

задает отображение виртуального документа `doc:VsdxDocument` на два компонента VSDX-документа `doc456.vdx`, находящегося в папке `vsdx`. Отображение интерпретирует целевой документ как ZIP-архив. Компонент `ent:Master1` указывает в архиве на файл мастеров `master1.xml`, а компонент `ent:Page1` — на разметку первой страницы документа `page1.xml`.

Обработка файлов XML выполняется с помощью объектов обработки DOM (Document Object Model), предоставляющих стандартизованный интерфейс доступа к XML-дереву. Загрузка в DOM-объект документа в формате единого XML-файла выполняется с помощью, например, такой HSM-директивы:

```
<dom:VdxProc srcDoc = "VdxDocument">
    <rcv:VdxExtract method="xslt" saveTo="dom:Extr" />
</dom:VdxProc>
```

Здесь создается DOM-объект `dom:VdxProc` и в него загружается XML-файл из виртуального документа `doc:VdxDocument`. заданного ранее. Далее с помощью приемника `rcv:VdxExtract` выполняется обработка загруженного файла и сохранение результата в другом DOM-объекте — `dom:Extr`. Атрибут `method="xslt"` указывает на то, что обработка (извлечение) выполняется путем XSL-трансформации на основе таблицы

стилей VdxExtract.xsl. При использовании формата на основе ZIP-архива HSM-директива может быть, например, такой:

```
<dom:VsdxProc srcDoc = "VsdxDocument.Page1">
    <src:Master1 srcDoc = "VsdxDocument.Master1"/>
    <rcv:VsdxExtract method="xslt" saveTo="dom:Extr"/>
</dom:VsdxProc>
```

Здесь в DOM-объект загружается XML-разметка первой страницы документа. После этого к ней с помощью источника src:Master1 подгружается XML-файл первого мастера, необходимый для выполнения извлечения.

Таким образом, сам процесс извлечения в иерархической модели не детализируется, указывается лишь метод его выполнения и ссылка на выполняющий преобразование компонент (в рассмотренном примере — на XSL-трансформатор). Компонент, выполняющий извлечение и построение концептуальной модели (в соответствии с рассмотренным выше алгоритмом, см. листинг 1, рис. 3), может быть реализован в виде таблицы стилей XSL-трансформации XML-документа (декларативный подход) или в виде модуля обработки, выполняющего манипуляции с деревом XML-документа через программный интерфейс DOM (процедурный подход).

Достоинства и недостатки подхода. Предложенный подход к извлечению семантической информации из графических документов обладает следующими *преимуществами*:

• Подход дает новую функциональность — возможность программным путем извлекать семантику схемы из ее графического представления, то есть формировать граф, отражающий элементы, соединения и существенные параметры схемы. Это может служить основой для автоматизированной проверки корректности схемы, генерирования технической документации и др.

• Подход применим к различным открытым графическим форматам на основе XML, которые предоставляют возможность программного доступа к внутреннему содержимому графических документов. Такие возможности имеются у большинства форматов, используемых в Visio-подобных графических редакторах.

• Экстракция схем не требует использования «родного» графического редактора, в среде которого получен графический документ. Эта особенность устраняет необходимость запуска графических редакторов или подключения их библиотек доступа на веб-сервере для программной обработки документов в веб.

• Простота реализации на основе ситуационно-ориентированной парадигмы. Реализованная в этой парадигме концепция виртуальных документов позволяет единообразно обрабатывать векторную графику в различных графических форматах и на этой основе генерировать результирующие документы — тоже в различных форматах.

Необходимо также отметить *ограничения и сложности* предложенного подхода. Данный подход применим только к открытым графическим форматам на основе XML, что снижает его универсальность. Другое ограничение в плане универсальности связано с допущениями об идентификации элементов схемы на основе трафаретов: подход работает, если при составлении схемы разработчик использовал для условленных графических обозначений заранее подготовленный набор фигур-трафаретов. Следует также отметить «хакерский» (в положительном смысле) оттенок подхода, затрудняющий его применение неквалифицированными пользователями. А именно, разработчик должен предварительно выполнить исследование внутренней структуры графической XML-разметки: во-первых, используемого графического формата, чтобы знать, как на внутреннем уровне задаются и идентифицируются фигуры и соединения графического изображения; во-вторых, набора трафаретов условных графических обозначений, используемых для построения схемы, чтобы знать, как на внутреннем уровне задаются их идентификационные и опциональные параметры. Только на базе этого можно правильно построить XPath-выражения для доступа к внутренним данным, которые необходимы для корректной работы алгоритма обработки графического документа и построения концептуальной модели схемы.

8. Практическое использование результатов. Рассмотренный выше подход был опробован и получил практическое применение в учебном процессе в Уфимском государственном авиационном техническом университете для извлечения свойств графических моделей в ходе курсового проектирования.

На рисунке 11 в качестве примера приведен фрагмент графического документа, содержащего схему — реляционную модель базы данных, которую студенты строят в среде графического редактора Visio на одном из этапов курсового проекта по дисциплине «базы данных». Здесь модель-диаграмма представлена в «уфимской» нотации: имена отношений (таблиц) представлены в прямоугольниках; имена атрибутов (столбцов) прикреплены к таблицам выносными линиями; концевые символы выносных линий задают свойства атрибутов; связи ссылочной целостности обозначены треугольниками, соединяющими атрибуты первичного

ключа родителя с атрибутами внешнего ключа ребенка. SQL-код создания соответствующих таблиц базы данных размещен в прямоугольных выносках, прикрепленных к таблицам пунктирными линиями. Метаданные документа представлены в «основной надписи», выполненной в соответствии со стандартом оформления конструкторской документации. Студентам-разработчикам предоставляется персонализированный бланк документа и набор фигур-трафаретов, содержащих фигуры таблиц, атрибутов, связей ссылочной целостности, выноски для записи SQL-кода. При составлении схемы на бланк переносятся копии трафаретов, соединяются между собой, вводятся имена таблиц и атрибутов, а также текст SQL-кода.

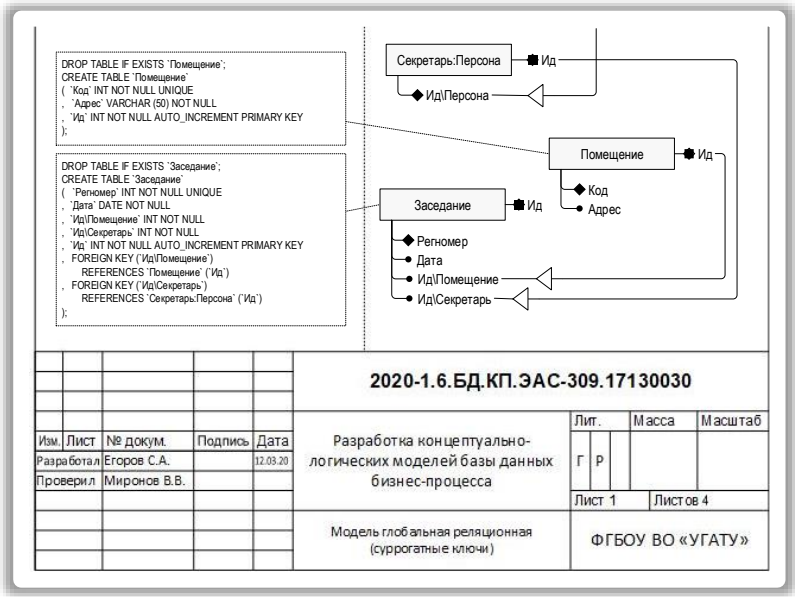


Рис. 11. Фрагмент примера схемы — реляционной модели базы данных: основная надпись документа (*внизу*); диаграмма реляционной модели (*справа*); SQL-код создания таблиц (*слева*)

В зависимости от решаемой задачи выполняется извлечение из этого графического документа следующих сведений:

• структуры реляционной модели — используется для автоматической проверки ее синтаксической корректности, а также соответствия заданию и результатам предшествующих этапов разработки;

• текста SQL-кода — используется для автоматической проверки соответствия SQL-кода, составленного студентом, диаграмме реляционной модели, а также для автоматической генерации заготовок проектной документации (например, документа «текст программы»);

• метаданных из основной надписи документа — для использования в отчетах и проектной документации. В результате достигается заметное уменьшение рутинных действий по контролю правильности составленной схемы и подготовки отчетной документации как у преподавателей, так и у студентов.

Таким образом, практическое применение предложенного подхода дает положительный эффект в виде нового качества за счет автоматизированного извлечения семантической информации из графических документов.

9. Заключение. В данной статье изложен научно обоснованный подход к извлечению семантической информации (знаний) из исходных графических данных. Этот подход обладает новизной, поскольку впервые позволяет в автоматизированном режиме извлекать структурно-параметрические свойства схем из электронных документов векторной графики, выполненных в открытых графических форматах на базе XML. Подход предполагает использование наборов заранее разработанных трафаретов для условных обозначений схемы. Предложенный в рамках подхода алгоритм построения концептуально-логического отображения предусматривает перебор фигур внутреннего представления графического документа и идентификацию условных графических обозначений, их соединений и свойств. В плане практической реализации подхода разработаны модели идентификации условных графических обозначений применительно к различным графическим форматам. Практическая реализация выполнена на основе ситуационно-ориентированной парадигмы, предусматривающей отображение виртуальных документов на графические файлы. Подход прошел успешную проверку работоспособности и получил практическое применение.

Литература

1. *Pieris D., Wijegunesekera M.C., Dias N.G.J.* ER model partitioning: Towards trustworthy automated systems development // International Journal of Advanced Computer Science and Applications. 2020. vol. 11, № 6. pp. 286–293. DOI: 10.14569/IJACSA.2020.0110638.
2. *Pérez R., Guerrero R.* A computer agent that develops visual compositions based on the ER-model // Annals of Mathematics and Artificial Intelligence. 2020. vol. 88, no. 5–6. P. 549–588. DOI: 10.1007/s10472-019-9616-3.
3. *Coelho D., Mueller K.* Infomages: Embedding Data into Thematic Images // Computer Graphics Forum. 2020. Vol. 39, no. 3. pp. 593–606. DOI: 10.1111/cgf.14004.

4. *Tsandilas T.* StructGraphics: Flexible Visualization Design through Data-Agnostic and Reusable Graphical Structures // IEEE Transactions on Visualization and Computer Graphics. 2021. vol. 27, no. 2. pp. 315–325. DOI: https://doi.org/10.1109/TVCG.2020.3030476.
5. *Yu Z., Xiong Z.* Comparative Analyses for the Performance of Rational Rose and Visio in Software Engineering Teaching // Journal of Physics: Conference Series. IOP Publishing. 2018. vol. 1087. no. 6. pp. 062–041. DOI: https://doi.org/10.1088/1742-6596/1087/6/062041.
6. *Parker D. J.* Mastering Data Visualization with Microsoft Visio Professional 2016 // Packt Publishing Ltd. 2016. P. 334.
7. *He L., Lian J.* Instructional Design of Practice Course of Logistics System Planning and Design Based on Visio // The 9th International Conference on Information Technology in Medicine and Education (ITME'2018), IEEE, 2018. pp. 526–530. DOI: 10.1109/ITME.2018.00122.
8. *Ruiz Ledesma E.F. et al.* Educational tool for generation and analysis of multidimensional modeling on data warehouse // International Journal of Advanced Computer Science and Applications. 2020. vol. 11, no. 9. pp. 261–267. DOI: 10.14569/IJACSA.2020.0110930.
9. *Shafiee S. et al.* Evaluating the benefits of a computer-aided software engineering tool to develop and document product configuration systems // Computers in Industry. 2021. vol. 128. DOI: 10.1016/j.compind.2021.103432.
10. *Medoh C., Telukdarie A.* Business Process Modelling Tool Selection: A review // IEEE International Conference on Industrial Engineering and Engineering Management (IEEM'2017). IEEE. 2017. pp. 524–528. DOI: 10.1109/IEEM.2017.8289946.
11. *Afanasyev A., Voit N., Gaynullin R.* The analysis of diagrammatic of workflows in design of the automated systems // Uncertainty Modelling in Knowledge Engineering and Decision Making. 2016. pp. 509-514. DOI: 10.1142/ 9789813146976_0082.
12. *Voit N., Bochkov S., Kirillov S.* Temporal Automaton RVTI-Grammar for the Diagrammatic Design Workflow Models Analysis // IEEE 14th International Conference on Application of Information and Communication Technologies (AICT'2020), Tashkent: Uzbekistan, 2020, pp. 1-6. DOI: 10.1109/AICT50176.2020.9368810.
13. *Afanasyev A., Voit N., Ukhanova M., Ionova I.* Development of the approach to check the correctness of workflows // Data Science and Knowledge Engineering for Sensing Decision Support (ITIDS'2018), pp. 1392-1399. DOI: 10.1142/9789813273238_0173.
14. *Shah R., Kesan J.* Interoperability challenges for open standards: ODF and OOXML as examples // Proceedings of the 10th Annual International Conference on Digital Government Research: Social Networks: Making Connections between Citizens, Data and Government (dg.o'09). Puebla: Digital Government Society of North America. 2009. pp. 56–62.
15. *Doncevic J., Fertalj K.* Database integration systems // Proceedings of 43rd International Convention on Information, Communication and Electronic Technology, (MIPRO'2020), 2020. pp. 1617–1622. DOI: .https://doi.org/10.23919/MIPRO48935.2020.9245245.
16. *Kolonko M., Mullenbach S.* Polyglot Persistence in Conceptual Modeling for Information Analysis // Proceedings of 10th International Conference on Advanced Computer Information Technologies, (ACIT'2020), 2020. pp. 590–594. DOI: https://doi.org/10.1109/ACIT49673.2020.9208928.
17. *Kosmerl I., Rabuzin K., Sestak M.* Multi-model databases - Introducing polyglot persistence in the big data world // Proceedings of 43rd International Convention on Information, Communication and Electronic Technology, (MIPRO'2020), 2020. pp. 1724–1729. DOI: 10.23919/MIPRO48935.2020.9245178.

18. *Montgomery C., Isah H., Zulkernine F.* Towards a Natural Language Query Processing System // Proceedings of 1st International Conference on Big Data Analytics and Practices (IBDAP'2020), 2020. DOI: 10.1109/IBDAP50342.2020.9245462.
19. *Миронов В.В., Гусаренко А.С., Юсупова Н.И.* Структурирование виртуальных мультидокументов в ситуационно-ориентированных базах данных с помощью entry-элементов // Труды СПИИРАН. 2017. № 53. С. 225–243. DOI: 10.15622/sp.53.11.
20. *Mironov V.V., Gusarenko A.S., Yusupova N.I.* Situation-oriented databases: document management on the base of embedded dynamic model // CEUR Workshop Proceedings (CEUR-WS.org): Selected Papers of the XI International Scientific-Practical Conference Modern Information Technologies and IT-Education (SITITO'2016), Moscow: Russia. 2016. vol. 1761. 2016. pp. 238-247.
21. *Mironov V., Gusarenko A., Yusupova N.* JSON Documents Processing Using Situation-Oriented Databases // Acta Polytechnica Hungarica. 2020. vol. 17. No. 8. pp. 29–40. DOI: 10.12700/APH.17.8.2020.8.3.
22. *Mironov V., Gusarenko A., Tuguzbaev G.* Graphic Documents Parametric Personalization for Information Support of Educational Design Using Situation-Oriented Databases // Advances in Intelligent Systems Research – Proceedings of the 8th Scientific Conference on Information Technologies for Intelligent Decision-Making Support (ITIDS'2020). pp. 260–267. DOI: 10.2991/aisr.k.201029.050

Миронов Валерий Викторович — д-р техн. наук, профессор, кафедра автоматизированных систем управления, факультет информатики и робототехники, УГАТУ. Область научных интересов: иерархическое моделирование, динамические модели, разработка баз данных и обработка данных в NoSQL СУБД. Число научных публикаций — 100. mironov@ugatu.su, http://hsm.ugatu.su; УГАТУ, ул. К. Маркса, 12, Уфа, 450008, РФ; р. т.: +7(347)273-77-17.

Гусаренко Артем Сергеевич — канд. техн. наук, доцент, кафедра автоматизированных систем управления, факультет информатики и робототехники, УГАТУ. Область научных интересов: иерархическое моделирование, динамические модели, разработка баз данных и обработка данных в NoSQL СУБД. Число научных публикаций — 50. gusarenko@ugatu.su, http://hsm.ugatu.su; http://itids.ugatu.su; http://csit.ugatu.su; УГАТУ, ул. К. Маркса, 12, Уфа, 450008, РФ; р. т.: +7(347)273-77-17.

Тугузбаев Гаяз Ахтямович — аспирант, кафедра автоматизированных систем управления, факультет информатики и робототехники, УГАТУ. Область научных интересов: разработка баз данных и обработка данных в NoSQL. Число научных публикаций — 3. tuguzbaev.g@ugatu.su; Карла Маркса, 12, Уфа, 450008, РФ; р. т.: +7(347)273-77-17.

Поддержка исследований. Работа выполнена при финансовой поддержке РФФИ (проект № 19-07-00682-а).

DOI 10.15622/ia.20.4.7

V. Mironov, A. Gusarenko, G. Tuguzbaev

EXTRACTING SEMANTIC INFORMATION FROM GRAPHIC SCHEMES

Mironov V., Gusarenko A., Tuguzbaev G. **Extracting Semantic Information from Graphic Schemes.**

Abstract. The problem of extracting semantic information from an electronic document specified in the vector graphics format and containing a graphic model (diagram) built using a graphic editor is considered. The problem is to program retrieving certain structural properties and parametric circuit and entering them into a database for later use. Based on the analysis of the capabilities of graphic editors, a conclusion has made about the relevance of this task for universal editors that are not tied to specific graphic notations and use open graphic document formats, which allows program processing. The proposed approach considers graphic documents at three levels of abstraction: conceptual (semantic properties of a schema), logical (presentation of semantic properties at the internal level of the document) and physical (internal organization of a graphic document). The solution to the problem is based on the construction of a conceptual-logical mapping, i.e., mapping a conceptual model of a circuit to a logical model of a graphic document, according to its physical model. Within the framework of the approach, an algorithm for constructing the indicated mapping is developed, presented in the form of an object-oriented pseudocode. The study of internal markup in open graphic formats made it possible to build models for identifying circuit elements and their connections to each other, which is necessary for a specific application of the algorithm. Expressions for addressing schema elements and accessing their properties are obtained. The proposed approach is implemented on the base of a situation-oriented paradigm, within which the extraction process is driven by a hierarchical situational model. The processed data is specified in the situational model in the form of virtual documents displayed on heterogeneous external data sources. For the problem being solved, we consider the mapping to two variants of vector graphics formats: to a "flat" markup file and to a set of such files in an electronic archive. The practical use of the results is illustrated by the example of extracting semantic information from graphical models developed at various stages of database design.

Keywords: Block Diagram, Vector Graphics, Property Extraction, Situational Paradigm, Hierarchical Situational Model, Virtual Document.

Mironov Valeriy — Dr. Tech. Sci., Professor, Department of Automated Control Systems, Faculty of Informatics and Robotics, USATU. Research interests: hierarchical modeling, dynamic models, database development and data processing in NoSQL DBMS. The number of scientific publications - 100. mironov@ugatu.su, http://hsm.ugatu.su; Karl Marx, 12, Ufa, 450008, Russian Federation; office phone: +7 (347) 273-77-17.

Gusarenko Artem — Candidate of Technical Sciences, Associate Professor, Department of Automated Control Systems, Faculty of Informatics and Robotics, USATU. Research interests: hierarchical modeling, dynamic models, database development and data processing in NoSQL DBMS. The number of scientific publications - 50. gusarenko@ugatu.su, http://hsm.ugatu.su; http://itids.ugatu.su; http://csit.ugatu.su; Karl Marx, 12, Ufa, 450008, Russian Federation; office phone +7 (347) 273-77-17.

Tuguzbaev Gayaz — Post-graduate student, Department of Automated Control Systems, Faculty of Informatics and Robotics, USATU. Research interests: object-oriented databases, DBMS,

database development and data processing in NoSQL. Number of scientific publications - 3. tuguzbaev.g@ugatu.su; Karl Marx, 12, Ufa, 450008, RF; office phone: +7 (347) 273-77-17.

Acknowledgements. This research is supported by RFBR (grant 19-07-00682-a).

References

1. Pieris D., Wijegunesekera M.C., Dias N.G.J. ER model partitioning: Towards trustworthy automated systems development // *International Journal of Advanced Computer Science and Applications*. 2020. vol. 11, № 6. pp. 286–293. DOI: https://doi.org/10.14569/IJACSA.2020.0110638.
2. Pérez R., Guerrero R. A computer agent that develops visual compositions based on the ER-model // *Annals of Mathematics and Artificial Intelligence*. 2020. vol. 88, no. 5–6. P. 549–588. DOI: https://doi.org/10.1007/s10472-019-9616-3.
3. Coelho D., Mueller K. Infomages: Embedding Data into Thematic Images // *Computer Graphics Forum*. 2020. vol. 39, no. 3. pp. 593–606. DOI: 10.1111/cgf.14004.
4. Tsandilas T. StructGraphics: Flexible Visualization Design through Data-Agnostic and Reusable Graphical Structures // *IEEE Transactions on Visualization and Computer Graphics*. 2021. vol. 27, no. 2. pp. 315–325. DOI: https://doi.org/10.1109/TVCG.2020.3030476.
5. Yu Z., Xiong Z. Comparative Analyses for the Performance of Rational Rose and Visio in Software Engineering Teaching // *Journal of Physics: Conference Series. IOP Publishing*. 2018. vol. 1087. no. 6. pp. 062–041. DOI: https://doi.org/10.1088/1742-6596/1087/6/062041.
6. Parker D. J. *Mastering Data Visualization with Microsoft Visio Professional 2016* // Packt Publishing Ltd. 2016. P. 334.
7. He L., Lian J. Instructional Design of Practice Course of Logistics System Planning and Design Based on Visio // The 9th International Conference on Information Technology in Medicine and Education (ITME'2018), IEEE, 2018. pp. 526–530. 10.1109/ITME.2018.00122.
8. Ruiz Ledesma E.F. et al. Educational tool for generation and analysis of multidimensional modeling on data warehouse // *International Journal of Advanced Computer Science and Applications.* 2020. vol. 11, no. 9. pp. 261–267. DOI: 10.14569/IJACSA.2020.0110930.
9. Shafiee S. et al. Evaluating the benefits of a computer-aided software engineering tool to develop and document product configuration systems // *Computers in Industry*. 2021. vol. 128. DOI: 10.1016/j.compind.2021.103432.
10. Medoh C., Telukdarie A. Business Process Modelling Tool Selection: A review // IEEE International Conference on Industrial Engineering and Engineering Management (IEEM'2017). IEEE. 2017. pp. 524–528. DOI: 10.1109/IEEM.2017.8289946.
11. Afanasyev A., Voit N., Gaynullin R. The analysis of diagrammatic of workflows in design of the automated systems // *Uncertainty Modelling in Knowledge Engineering and Decision Making*. 2016. pp. 509-514. DOI: 10.1142/9789813146976_0082.
12. Voit N., Bochkov S., Kirillov S. Temporal Automaton RVTI-Grammar for the Diagrammatic Design Workflow Models Analysis // IEEE 14th International Conference on Application of Information and Communication Technologies (AICT'2020), Tashkent: Uzbekistan, 2020, pp. 1-6. DOI: 10.1109/AICT50176.2020.9368810.
13. Afanasyev A., Voit N., Ukhanova M., Ionova I. Development of the approach to check the correctness of workflows // Data Science and Knowledge Engineering for Sensing Decision Support (ITIDS'2018), pp. 1392-1399. DOI: 10.1142/9789813273238_0173.
14. Shah R., Kesan J. Interoperability challenges for open standards: ODF and OOXML as examples // Proceedings of the 10th Annual International Conference on Digital Government Research: Social Networks: Making Connections between Citizens, Data and

Government (dg.o'09). Puebla: Digital Government Society of North America. 2009. pp. 56–62.

15. Doncevic J., Fertalj K. Database integration systems // Proceedings of 43rd International Convention on Information, Communication and Electronic Technology, (MIPRO'2020), 2020. pp. 1617–1622. DOI: 10.23919/MIPRO48935.2020.9245245.
16. Kolonko M., Mullenbach S. Polyglot Persistence in Conceptual Modeling for Information Analysis // Proceedings of 10th International Conference on Advanced Computer Information Technologies, (ACIT'2020), 2020. pp. 590–594. DOI: 10.1109/ACIT49673.2020.9208928.
17. Kosmerl I., Rabuzin K., Sestak M. Multi-model databases - Introducing polyglot persistence in the big data world // Proceedings of 43rd International Convention on Information, Communication and Electronic Technology, (MIPRO'2020), 2020. pp. 1724–1729. DOI: 10.23919/MIPRO48935.2020.9245178.
18. Montgomery C., Isah H., Zulkernine F. Towards a Natural Language Query Processing System // Proceedings of 1st International Conference on Big Data Analytics and Practices (IBDAP'2020), 2020. DOI: 10.1109/IBDAP50342.2020.9245462.
19. Mironov V.V., Gusarenko A.S., Yusupova N.I. Structuring virtual multi-documents in situationally oriented databases by means of entry-elements // *SPIIRAS Proceedings*. 2017. vol. 4, № 53. pp. 225–242. DOI: 10.15622/sp.53.11.
20. Mironov V.V., Gusarenko A.S., Yusupova N.I. Situation-oriented databases: document management on the base of embedded dynamic model // CEUR Workshop Proceedings (CEUR-WS.org): Selected Papers of the XI International Scientific-Practical Conference Modern Information Technologies and IT-Education (SITITO'2016), Moscow: Russia. 2016. vol. 1761. 2016. pp. 238-247.
21. Mironov V., Gusarenko A., Yusupova N. JSON Documents Processing Using Situation-Oriented Databases // *Acta Polytechnica Hungarica*. 2020. vol. 17. no. 8. pp. 29–40. DOI: 10.12700/APH.17.8.2020.8.3.
22. Mironov V., Gusarenko A., Tuguzbaev G. Graphic Documents Parametric Personalization for Information Support of Educational Design Using Situation-Oriented Databases // Advances in Intelligent Systems Research – Proceedings of the 8th Scientific Conference on Information Technologies for Intelligent Decision-Making Support (ITIDS'2020). pp. 260–267. DOI: 10.2991/aisr.k.201029.050.

УДК 519.682.3, 519.245, 519.644.7 **DOI** 10.15622/ia.20.4.8

С.М. Абрамов, В.А. Роганов, В.И. Осипов, Г.А. Матвеев

РЕАЛИЗАЦИЯ ПАКЕТА LAMMPS НА Т-СИСТЕМЕ С ОТКРЫТОЙ АРХИТЕКТУРОЙ

С.М. Абрамов, В.А. Роганов, В.И. Осипов, Г.А. Матвеев **Реализация пакета LAMMPS на Т-системе с открытой архитектурой.**

Аннотация. Суперкомпьютерные приложения обычно реализуются на языках программирования C, C++, Fortran с использованием различных вариантов библиотеки Message Passing Interface. В проекте "Т-система" (OpenTS) исследуются вопросы автоматического динамического распараллеливания программ. С практической точки зрения актуальна реализация приложений в смешанном (гибридном) стиле, когда часть приложения пишется в парадигме автоматического динамического распараллеливания программ и не использует никаких примитивов библиотеки MPI, а другая его часть пишется с использованием библиотеки Message Passing Interface.

В этом случае используется библиотека, которая входит в состав Т-системы и имеет название DMPI (Dynamic Message Passing Interface). Необходимо оценить эффективность реализации MPI, которая есть в Т-системе. Целью данной работы является исследование эффективности реализации DMPI в Т-системе. В классическом MPI приложении 0% кода реализовано с помощью автоматического динамического распараллеливания программ и 100% кода реализовано в виде обычной Message Passing Interface программы.

Для сравнительного анализа в начале код выполняется на стандартном Message Passing Interface, для которого он был написан изначально, и потом этот код выполняется с использованием библиотеки DMPI, входящей в состав Т-системы. При сравнении эффективности подходов оцениваются потери производительности и перспективность применения гибридного стиля программирования. В результате проведенных экспериментальных исследований для разных типов вычислительных задач удалось убедиться, что потери эффективности пренебрежимо малы. Это позволило сформулировать направление дальнейшей работы над Т-системой и наиболее перспективные варианты построения гибридных приложений.

В настоящей статье приводятся результаты сравнительных испытаний приложения LAMMPS с использованием OpenMPI и с использованием OpenTS DMPI. Результаты испытаний подтверждают эффективность реализации DMPI в среде параллельного программирования OpenTS.

Ключевые слова: параллельный алгоритм, язык программирования T++, OpenTS, Т-система, молекулярная динамика, перидинамика, Т-приложение, benchmark.

1. Введение. Работа над реализациями протокола MPI (Message Passing Interface) ведется как в коммерческих, так и в научных лабораториях. Протокол MPI был разработан группой, в которую входили сотрудники лабораторий коммерческих компаний: Аргоннской национальной лаборатории и Университета штата Миссисипи. Сегодня доступны следующие свободные реализации MPI: MPICH, MVAPICH и OpenMPI.

Аргоннская национальная лаборатория продолжает заниматься проектом MPICH при финансовой поддержке правительства США [1].

MVAPICH разработан Университетом штата Огайо [2]. Крупные производители параллельных вычислительных систем имеют свои реализации MPI: Cray MPI [3], Tianhe MPI [4], Intel MPI [5], IBM Blue Gene/Q MPI [6], IBM PE MPICH [7], IBM Platform MPI [8], SGI MPI [9], Fujitsu MPI [10], MS MPI [11]. В настоящее время версия протокола MPI-3 поддерживается всеми основными реализациями MPI. Ведется разработка протокола MPI-4 [12]. MPI является одним из важных средств поддержки параллельных вычислений. Большинство суперкомпьютерных приложений опирается на MPI. От эффективности реализации MPI зависит эффективность и масштабируемость реализации приложений.

OpenTS [13-20] – система для параллельного программирования, поддерживающая динамически загружаемые адаптеры для коммуникационного уровня. В системе OpenTS реализован язык для параллельных вычислений T++, который является расширением языка программирования C++. Синтаксис языка T++ отличается от синтаксиса языка C++ добавлением в него нескольких ключевых слов.

Система OpenTS использует собственную реализацию MPI, оформленную в виде динамической библиотеки OpenTS DMPI (Dynamic MPI). Библиотека OpenTS DMPI предоставляет базовое подмножество функций из стандарта MPI либо за счет переадресации вызовов к локальной библиотеке MPI, установленной на целевой системе, либо поверх протокола TCP/IP. При инициализации системы OpenTS активируется подсистема DMPI. Эта подсистема динамически загружает ту локальную библиотеку MPI, которая указана в переменных окружения приложения. Некоторые из функций MPI реализованы разработчиками системы OpenTS напрямую. В OpenTS DMPI протокол MPI-2 реализован не полностью.

Язык параллельного программирования T++ разработан таким образом, что необходимости использования в нем MPI функций нет. Однако, если для достижения эффективности работы программы разработчик желает использовать MPI функции явно, то он может это сделать.

Язык T++ позволяет автоматически динамически распараллеливать приложения. Он был разработан для того, чтобы приложения писались комфортно и, вследствие, освобождали программиста от планирования процессов, распределения вычислительной нагрузки, передачи данных между процессами, а также их синхронизации. Если в приложении OpenTS выходить за рамки языка T++ и пользоваться прямыми передачами, которые реализует MPI, то есть часть работы выполнять на языке T++, а часть работы выполнять, опираясь на передачу данных

вручную, указывая те или другие пересылки данных между параллельными процессами, то это несущественно усложнит код, но может повысить эффективность.

Если это естественно, то этим следует пользоваться. Тогда код будет более прозрачным, управляемым, понятным. Это возможно, так как реализация OpenTS содержит в себе MPI – в приложении можно вызывать его функции.

Реализация протокола MPI в системе для параллельного программирования OpenTS называется DMPI. Суть подхода состоит в том, что пишется приложение на языке T++, но там, где это разумно, напрямую используются примитивы DMPI.

Решаемая в данной работе научная задача. Глобальная цель проекта OpenTS состоит в поддержке гибридной модели вычислений, когда часть приложения написана на T++, а в другой части вызываются функции MPI.

Прежде чем двигаться в сторону полной реализации глобальной цели исследования, правильно было бы ответить на вопрос: достаточно ли эффективна OpenTS DMPI? Задача, которая решается, – это показать, что использование такой гибридной модели не проигрывает в эффективности в сравнении с исходной реализацией приложения, а также имеет низкие накладные расходы, тем самым дает возможность для развития. Для этого используется известное приложение LAMMPS.

LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator) — пакет для классической молекулярной динамики, написанный группой из Сандийских национальных лабораторий [21]. Для работы на многопроцессорных системах используется интерфейс MPI. В статье описывается реализация симулятора молекулярной динамики LAMMPS с использованием библиотеки OpenTS DMPI и сравнивается эффективность полученной реализации с оригинальной версией, скомпонованной с локальной библиотекой OpenMPI. Данное сравнение производится с целью продемонстрировать низкие накладные расходы DMPI, а также возможность встраивания различных готовых параллельных решателей и симуляторов в среду динамического распараллеливания OpenTS для их совместного использования с алгоритмами интеллектуального поиска, например, для оптимизации структуры и состава новых, синтетических материалов или конструкций.

В процессе исследования было разработано и добавлено в модуль DMPI несколько MPI функций, а также проведено сравнительное тестирование с реализацией на OpenMPI. Новые MPI функции были добавлены в связи с тем, что они не были реализованы в предыдущей версии OpenTS DMPI.

В таблице 1 приводятся результаты сравнительных испытаний LAMMPS с использованием OpenMPI и OpenTS DMPI. Результаты испытаний подтверждают эффективность реализации протокола MPI в системе параллельного программирования OpenTS. На каждом этапе испытаний вычислялось значение средней производительности и проводилось сравнение её значений для двух платформ.

Положительное значение отклонения производительности означает, что производительность OpenMPI выше производительности OpenTS DMPI. Если значение отклонения производительности отрицательное, то производительность OpenTS DMPI выше производительности OpenMPI.

Таблица 1. Сравнение средних значений производительности LAMMPS OpenMPI и LAMMPS OpenTS DMPI

Пример LAMMPS	Минимальное отклонение средней производительности (%)	Максимальное отклонение средней производительности (%)	Среднее отклонение производительности (%)
accelerate	-2,7	0,2	-1,2
airebo	-1,3	1,1	-0,8
ASPHERE	-1,9	3,7	1
atm	1,5	3,3	2,5
balance	-1,1	2,6	-0,1
colloid	-1,4	3,1	0,5
comb	-2,3	-0,4	-1
crack	-3,2	1,9	-0,5
deposit	-1,4	0,5	-0,6
DIFFUSE	-0,9	1,2	0,4
ELASTIC-T	-1,7	4	0,7
flow	-3,8	-0,5	-1,5
friction	-1,3	0,9	-0,3
indent	-1,5	1,1	-0,2
KAPPA	-0,8	0,8	-0,2
min	-1,5	1,1	-0,5
nemd	-2,3	4	0,4
obstacle	-3,9	0,5	-1,1
peri	-4,9	5,6	0,3
USER diffraction	-4,3	-2,6	-3,5
USER dpd	-1,1	0,6	-0,2
VISCOSITY	-0,8	5,8	2,5

Каждый из разделов статьи является описанием одного из примеров LAMMPS. Список этих примеров следующий: accelerate, airebo,

ASPHERE, atm, balance, colloid, comb, crack, deposit, DIFFUSE, ELASTIC-T, flow, friction, indent, KAPPA, min, nemd, obstacle, peri, USER diffraction, USER dpd, viscosity. С каждым примером из списка было проведено по 10 испытаний. Каждое испытание проводилось на количестве ядер процессора от 1 до 8 с шагом 1. На каждом этапе испытаний вычислялось значение средней производительности. Испытания проведены для двух платформ: OpenMPI и OpenTS DMPI.

Цель состояла в проверке текущих показателей производительности (ухудшение/улучшение) при переносе примера с платформы OpenMPI на платформу OpenTS DMPI. При этом вычисляется производительность примера LAMMPS или выбранного шага примера LAMMPS. Это связано с тем, что, если программа LAMMPS состоит из нескольких шагов, то LAMMPS вычисляет производительность каждого шага и не вычисляет производительности всей программы.

2. Модель Гея-Берне для двухосных эллипсоидных мезогенов в изотропной фазе. Пример accelerate [23]. Модель Гея-Берне [24,25] широко используется при моделировании жидкокристаллических систем. Модель Гея-Берне является анизотропной формой потенциала Леннарда-Джонса [22]. Она описывает взаимодействие между частицами, имеющими форму эллипсоидов. Модель учитывает не только расстояние между центрами частиц, но и их ориентацию. Пример состоит из двух шагов. В начале для системы атомов задается определенная температура, для каждого атома случайным образом задается скорость. На первом этапе поддерживается постоянная температура, и постепенно увеличивается давление до заданного значения. На втором этапе поддерживается постоянный объем и энергия системы частиц. Было проведено сравнительное испытание производительности примера LAMMPS accelerate OpenMPI и LAMMPS accelerate OpenTS DMPI.

3. Полиэтилен с потенциалом межчастичного взаимодействия AIREBO. Пример airebo [23]. Потенциал AIREBO (Adaptive Intermolecular Reactive Empirical Bond Order) [26,27] состоит из трех составляющих:

$$E = \frac{1}{2}\sum_{i}\sum_{j\neq i}\left[E_{ij}^{\mathrm{REBO}} + E_{ij}^{\mathrm{LJ}} + \sum_{k\neq i,j}\sum_{l\neq i,j,k} E_{kijl}^{\mathrm{TORSION}}\right].$$

Первое слагаемое соответствует потенциалу Бреннера [28] второго поколения (REBO), который часто используется для моделирования взаимодействия между атомами углерода и водорода. Оно отвечает за короткодействующее взаимодействие, на расстоянии менее 2 ангстрем. Второе слагаемое отвечает за потенциал Леннарда-Джонса [22],

который действует на расстояниях от 2 ангстрем до величины радиуса экранирования. Третье слагаемое отвечает за определение углов между связями в конфигурациях углеводородов. Во входном файле описаны координаты атомов молекулы полиэтилена, состоящей из 20 атомов углерода и 40 атомов водорода, затем эта молекула реплицируется на систему из 32640 атомов. Задается начальная температура 300 градусов по Кельвину, случайным образом устанавливаются скорости атомов, затем в течение ста временных шагов поддерживается постоянный объем и энергия. Было проведено сравнительное испытание производительности примера LAMMPS airebo OpenMPI и LAMMPS airebo OpenTS DMPI.

4. Диффузия при использовании метода стохастической вращательной динамики (SRD) для жестких прямоугольных частиц. Пример ASPHERE [23]. При использовании метода стохастической вращательной динамики [29] частицы растворителя не взаимодействуют друг с другом, а взаимодействуют с частицами растворенного вещества, которые могут иметь разную форму, например, эллипсоиды, димеры, прямоугольники и т.д. [30, 31] (рис. 1). В частности, одной из характеристик SRD (Stochastic Rotation Dynamics) является средний путь до столкновения λ, который вычисляется по формуле:

$$\lambda = \Delta t_{SRD} \sqrt{\frac{k_B T_{SRD}}{m}},$$

где Δt_{SRD} – шаг времени, k_B – постоянная Больцмана, T_{SRD} – температура, m – масса частицы. В примере выполняется два этапа. На первом этапе создается 30 прямоугольных наночастиц, каждая из которых состоит из 14 атомов. Задается потенциал (только для первого этапа):

$$E = A\left[1 + \cos\left(\frac{\pi r}{r_c}\right)\right], \qquad r < r_c,$$

где r_c – радиус отсечки, A – переменная, которая в начале этапа принимает значение 0, затем линейно по времени увеличивается к концу этапа до 30. С помощью этого потенциала близко расположенные или перекрывающиеся частицы растаскиваются друг от друга.

На втором этапе пространство между частицами заполняется раствором, задается потенциал Леннарда-Джонса, параметры которого устанавливаются таким образом, что взаимодействия через потенциал

между частицами растворителя нет. Запускается команда fix srd, которая отрабатывает взаимодействие между наночастицами и частицами растворителя. Было проведено сравнительное испытание производительности примера LAMMPS ASPHERE OpenMPI и LAMMPS ASPHERE OpenTS DMPI.

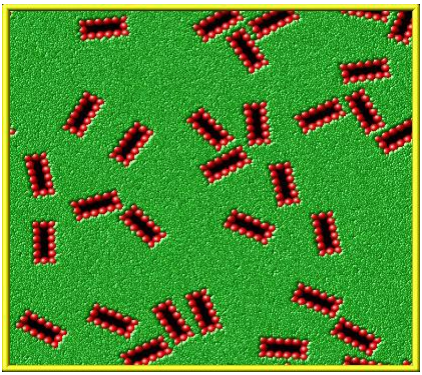


Рис. 1. Демонстрация метода стохастической вращательной динамики (SRD) для жестких прямоугольных частиц в растворителе

5. Трехчастичный потенциал Аксильрода–Теллера–Муто. Пример atm [23]. Трехчастичный потенциал вычисляется для троек взаимно близко расположенных частиц. Наиболее известный из трехчастичных потенциалов — потенциал Аксильрода–Теллера–Муто [32,33]. Если три частицы образуют треугольник с углами γ_1, γ_2 γ_3 и сторонами r_{12}, r_{23}, r_{31}, потенциал вычисляется по формуле

$$U_{123} = C_{123}\frac{1+3\cos\gamma_1\cos\gamma_2\cos\gamma_3}{r_{12}^3 r_{23}^3 r_{31}^3},$$

где $C_{123} = \frac{3\hbar}{\pi}\int_0^\infty \alpha_1(i\omega)\alpha_2(i\omega)\alpha_3(i\omega)d\omega$ [34],

α_1, α_2, α_3 – поляризуемость. В примере на атомы действуют одновременно 2 потенциала, Леннарда-Джонса и Аксильрода–Теллера–Муто при постоянном объеме и температуре. Было проведено сравнительное испытание производительности примера LAMMPS atm OpenMPI и LAMMPS atm OpenTS DMPI.

6. Балансировка. При балансировке область вычислительной системы, в которой производится симуляция, распределяются между

вычислительными узлами (процессорами, ядрами). Если во входном файле нет команд балансировки, то область симуляции распределяется между узлами с помощью грида (сетки) с равными ячейками (рис. 2).

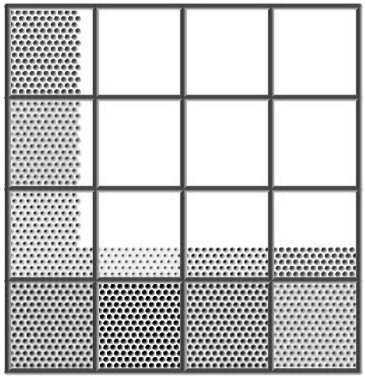
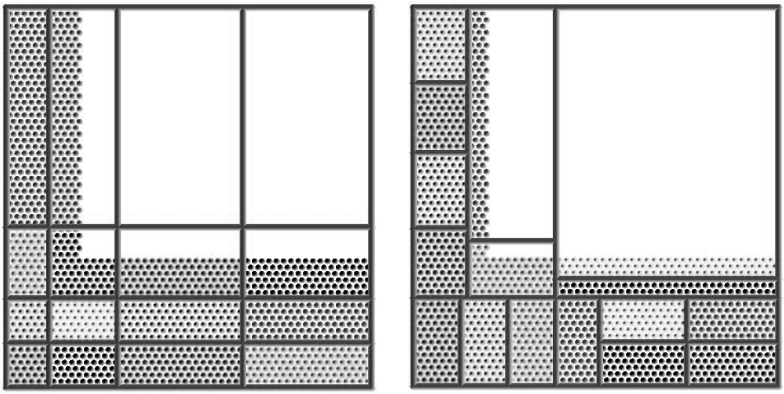


Рис. 2. Балансировка вычислительной системы, имеющей 16 процессоров в случае, когда нет команд балансировки

С помощью команды balance [35] область симуляции можно перераспределить один раз, например, в начале программы (рис. 3a). Область симуляции будет идеально сбалансирована, если ячейки разбиения имеют одинаковое количество частиц. Команда balance имеет два стиля, стиль shift (разбиение с помощью грида) и стиль rcb (плиточное разбиение) (рис. 3b).



a) b)

Рис. 3. Балансировка с помощью команды balance: a) стиль shift; b) стиль rcb

Команда balance производит балансировку один раз (статическая балансировка). Если балансировку следует проводить динамически несколько раз в течении работы программ, используют команду fix balance [36]. Визуализация симуляции представлена на рисунке 4.

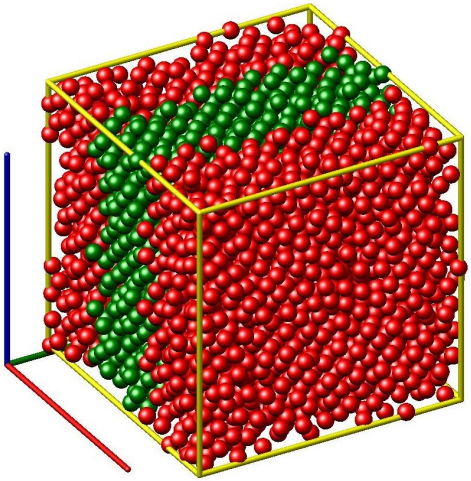


Рис. 4. Пример LAMMPS balance: статическая балансировка, стиль shift

Область симуляции является кубом. При выполнении программы куб по оси *x* режется на количество параллелепипедов, равное количеству процессоров. Каждая из подобластей закрепляется за определенным процессором. Поддерживается постоянный объем и энергия. Было проведено сравнительное испытание по оцениванию производительности примеров LAMMPS balance OpenMPI и LAMMPS balance OpenTS DMPI.

7. Коллоидный раствор. Пример colloid [23]. В примере описано 2 типа частиц, частицы растворителя, имеющие массу 1 и тяжелые частицы с массой 9 (рис. 5). Частицы растворителя взаимодействуют друг с другом по формуле Леннарда-Джонса. Формулы потенциалов взаимодействия частиц коллоида между собой и частиц коллоида и растворителя приведены в [59, 60]. Коллоидный раствор состоит из больших частиц растворенного вещества и маленьких частиц растворителя. При симуляции поддерживается постоянная температура. Давление постепенно увеличивается до заданного значения. Было проведено сравнительное испытание производительности примера LAMMPS colloid OpenMPI и LAMMPS colloid OpenTS DMPI.

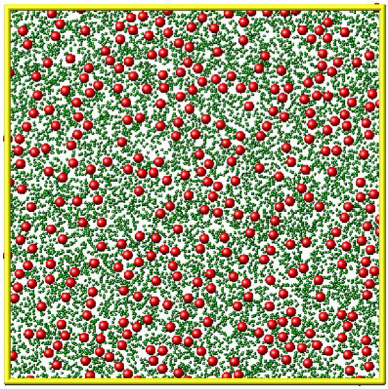


Рис. 5. Коллоидный раствор

8. Многочастичный потенциал COMB. При использовании потенциалов COMB (Charge-Optimized Many-Body) [37] и COMB3 [38] энергия системы из нескольких атомов имеет вид:

$$E_T = \sum_i [E_i^{self}(q_i) + \sum_{j>i} [E_{ij}^{short}(r_{ij}, q_i, q_j) + E_{ij}^{Coul}(r_{ij}, q_i, q_j)] +$$
$$+ E^{polar}(q_i, r_{ij}) + E^{vdW}(r_{ij}) + E^{barr}(q_i) + E^{corr}(r_{ij}, \theta_{jik})],$$

где E_i^{self} – энергия i-го атома, включая энергию ионизации атома и энергию сродства к электрону,

E_{ij}^{short} – BO-потенциал (Bond order potential) [39],

E_{ij}^{Coul} – кулоновское взаимодействие,

E^{polar} – поляризация (только для потенциала COMB3),

E^{vdW} – взаимодействие Ван-дер-Ваальса (только для потенциала COMB3),

E^{barr} – барьерная функция,

E^{corr} – угловая поправка.

Потенциал COMB использовался в примере LAMMPS comb [23]. Моноклинный оксид гафния был помещен в термостат Нозе-Хувера [40, 41]. Термостат используется для поддержания постоянной температуры в системе. Уравнения термостата:

$$\frac{dv}{dt} = \frac{F(t)}{m} - \zeta v(t),$$
$$\frac{d\zeta}{dt} = \frac{1}{Q}[\sum mv(t)^2 - (X+1)k_B T],$$

где Q – параметр, X – количество степеней свободы.

Термостат реализован в программе LAMMPS с помощью команды fix nvt [42]. Визуализация примера представлена на рисунке 6.

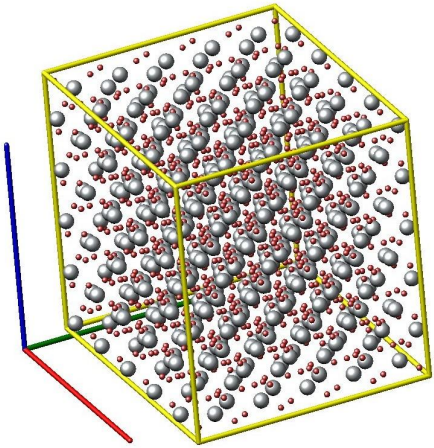


Рис. 6. Оксид гафния. Пример LAMMPS comb

Заранее заданные координаты 1500 атомов считываются из входного файла. Поддерживается постоянная температура 300 градусов по Кельвину. Было проведено сравнительное испытание производительности примера LAMMPS comb OpenMPI и LAMMPS comb OpenTS DMPI.

9. Разрыв листа из твердого материала. Пример crack [23]. Лист тянут вверх за верхнюю полоску, и он разрывается (рис. 7). Описывается 6 регионов области симуляции. Взаимодействие между левым верхним и левым нижним регионом отключают. Верхнюю часть тянут вверх с постоянной скоростью. Нижняя часть неподвижна. Скорость движения вверх частиц, находящихся между верхней и нижней частью, пропорциональна их ординате. Было проведено сравнительное испытание производительности примера LAMMPS crack OpenMPI и LAMMPS crack OpenTS DMPI.

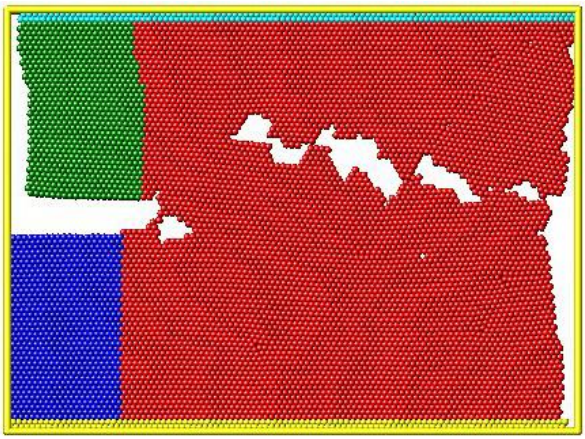


Рис. 7. Разрыв листа твердого материала

10. Нанесение молекул на поверхность. Пример deposit [23]. Гибкие димеры падают на подложку (рис. 8). Поддерживается постоянная энергия и объем. Было проведено сравнительное испытание производительности примера LAMMPS deposit OpenMPI и LAMMPS deposit OpenTS DMPI. Входной файл был модифицирован, увеличен период, с которым падают димеры сверху и количество временных шагов симуляции.

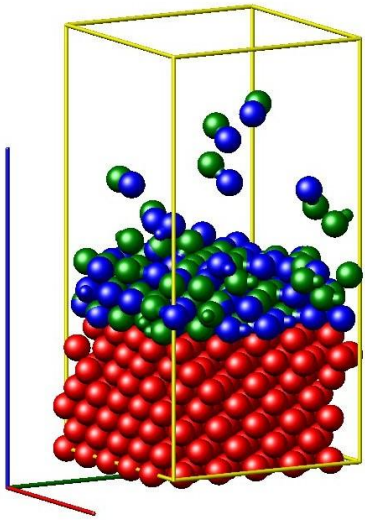


Рис. 8. Нанесение молекул на поверхность

11. Вычисление коэффициента диффузии методом среднеквадратичного смещения (Mean Suared Displacement). Пример DIFFUSE [23]. Обозначим через MSD(t) функцию:

$$MSD(t) = \frac{1}{N}\sum_{i=1}^{N} | x_i(t) - x_i(0) |^2,$$

где $x_i(t)$ – положение i-й частицы в момент времени t. Если предположить, что движение частиц является броуновским, и прошло достаточно много времени, можно считать, выполняется равенство [43]:

$$MSD(t) \approx 2nDt,$$

где n – размерность пространства, D – коэффициент диффузии. С помощью последнего равенства можно вычислить коэффициент диффузии (рис. 9). Функция MSD вычисляется в LAMMPS с помощью команды compute msd [44]. График функции MSD для примера представлен на рисунке 10.

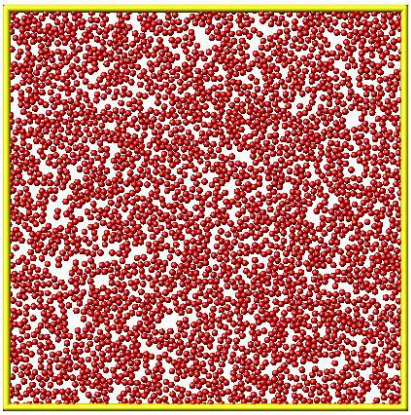


Рис. 94. Вычисление коэффициента диффузии

По горизонтальной оси откладывается время, по вертикальной – значение функции MSD. Чтобы вычислить среднеквадратичное смещение MSD, в примере используется команда compute msd. Было проведено сравнительное испытание производительности примера LAMMPS diffusion OpenMPI и LAMMPS diffusion OpenTS DMPI (рис. 10).

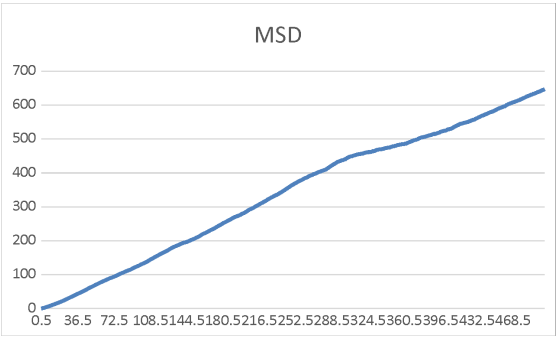


Рис. 10. График функции MSD

12. Вычисление упругих постоянных тензора упругости для кристалла. Пример ELASTIC-T [23]. Образец подвергается нагрузке, проводятся измерения, находятся упругие постоянные [45, 46]. Кристалл кремния, состоящий из 216 атомов, подвергается сжатию в разных направлениях, производится 6 испытаний. При сжатии координаты атомов меняются пропорционально величине сжатия. В примере используется трехчастичный потенциал Стиллингера-Вебера [47]. Было проведено сравнительное испытание производительности примера LAMMPS ELASTIC-T OpenMPI и LAMMPS ELASTIC-T OpenTS DMPI.

13. Течение Куэтта. Пример flow [23]. Для трехмерного пространства течение Куэтта [48] представляет собой течение между двумя параллельными стенками, когда одна из стенок движется параллельно другой стенке с постоянной скоростью. Среди примеров LAMMPS представлен 2-мерный вариант течения Куэтта, в этом случае течение жидкости происходит в канале с параллельными берегами, верхний берег движется параллельно другому с постоянной скоростью (рис. 11). Было проведено сравнительное испытание производительности примера LAMMPS flow OpenMPI и LAMMPS flow OpenTS DMPI. Входной файл был модифицирован, был увеличен размер области симуляции и, соответственно, количество частиц.

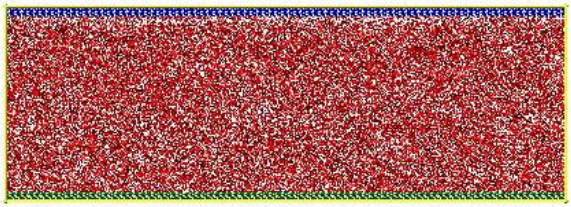
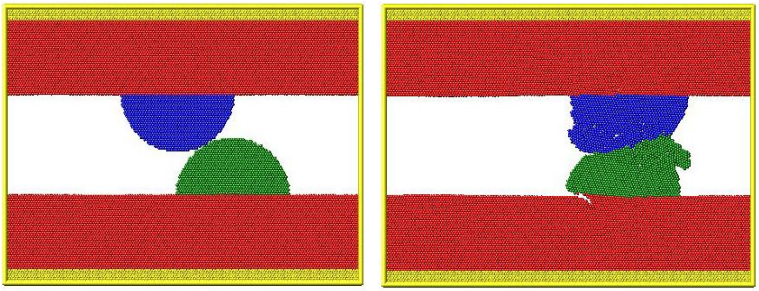


Рис. 115. Течение Куэтта. Верхний берег движется влево с постоянной скоростью

14. Фрикционный контакт полукруглых выпуклостей в двумерном пространстве. Пример friction [23]. Синяя полусфера движется в сторону зеленой (рис. 12a). Результат соприкосновения показан на рисунке 12b. Было проведено сравнительное испытание производительности примера LAMMPS friction OpenMPI и LAMMPS friction OpenTS DMPI. Входной файл был модифицирован, увеличен размер области симуляции и диаметр полусфер.



a) b)

Рис. 12. Трение двух полукруглых выпуклостей: a) начало соприкосновения; b) конец симуляции

15. Испытание с использованием сферического индентора. Пример indent [23]. Сферический индентор вдавливается в двумерный твердый образец, затем убирается (рис. 13).

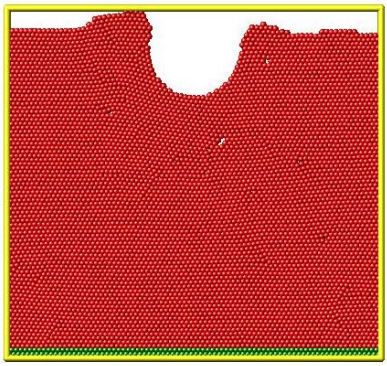


Рис. 6. Испытание с использованием сферического индентора

Для симулирования индентора используется команда fix indent.

Было проведено сравнительное испытание производительности примера LAMMPS indent OpenMPI и LAMMPS indent OpenTS DMPI. Входной файл был модифицирован, увеличен размер области симуляции и радиус индентора.

16. Вычисление коэффициента теплопроводности жидкости с потенциалом межчастичного взаимодействия Леннарда-Джонса. Пример KAPPA [23]. Согласно закону теплопроводности Фурье,

$$Q = -k\, grad(T),$$

где Q – тепловой поток,
k – коэффициент теплопроводности,
T – температура.

Эту формулу используют для определения коэффициента теплопроводности. В области симуляции, которая является параллелепипедом, описываются две подобласти: холодная и горячая. В холодной части поддерживается низкая температура, в горячей – высокая. Выбор холодной и горячей подобласти осуществляют таким образом, что градиент температуры направлен по оси z. Для поддержки высокой и низкой температуры в заданных областях используется команда fix heat [49]. Вычисление коэффициента теплопроводности осуществляется с использованием команды compute ke, которая рассчитывает кинетическую энергию системы частиц. Подробности реализации описаны в [50]. Было проведено сравнительное испытание производительности примера LAMMPS KAPPA OpenMPI и LAMMPS KAPPA OpenTS DMPI.

17. Минимизация энергии. Пример min [23]. При минимизации энергии целевой функцией является потенциальная функция:

$$E(r_1, r_2, \ldots, r_N) = \sum_{i,j} E_{pair}(r_i, r_j) + \sum_{ij} E_{bond}(r_i, r_j) +$$
$$+ \sum_{ijk} E_{angle}(r_i, r_j, r_k) + \sum_{ijkl} E_{dihedral}(r_i, r_j, r_k, r_l) +$$
$$+ \sum_{ijkl} E_{improper}(r_i, r_j, r_k, r_l) + \sum_{i} E_{fix}(r_i),$$

где E_{pair} – парное взаимодействие частиц, включая кулоновское,

E_{bond} – энергия взаимодействия между определенными парами частиц. Список таких частиц определяется во входном файле,

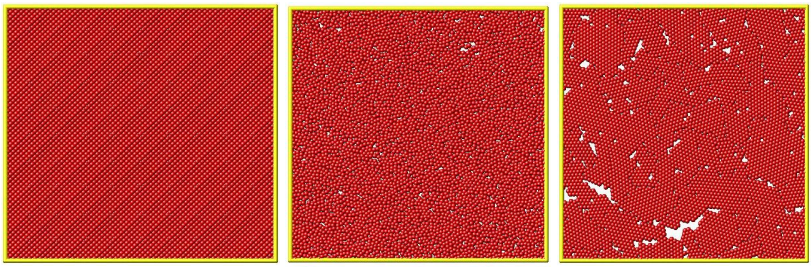
E_{angle} – энергия взаимодействия троек атомов,

$E_{dihedral}$ – энергия взаимодействия четверок атомов в случае правильных торсионных углов,

$E_{improper}$ – энергия взаимодействия атомов в случае неправильных торсионных углов,

E_{fix} – поправка для случая дополнительных ограничений.

В процессе симуляции вектора скоростей атомов меняются таким образом, чтобы потенциальная функция уменьшалась, пока не достигнет локального минимума. Симуляция проводилась в два этапа. Начальное расположение частиц расплава с потенциалом межчастичного взаимодействия Леннарда-Джонса представлено на рисунке 14a. После того, как система поддерживалась при постоянном объеме и энергии заданное количество временных шагов, частицы перемешались (рис. 14b). После минимизации был достигнут локальный минимум потенциальной функции (рис. 14c). Минимизация энергии производится с помощью команды minimize. Было проведено сравнительное испытание производительности примера LAMMPS min OpenMPI и LAMMPS min OpenTS DMPI. Входной файл был модифицирован, увеличен размер области симуляции по сравнению с исходным примером.



a) b) c)

Рис. 14. Минимизация энергии: a) начало; b) первый этап; c) достигнут локальный минимум потенциальной функции

18. Неравновесная молекулярная динамика. Неравновесная молекулярная динамика [51] описывает модели неравновесных систем. Например, это могут быть молекулярные модели, в которых на систему частиц оказывается внешнее воздействие. Один из примеров неравновесной молекулярной динамики рассматривался ранее.

19. Вычисление коэффициента теплопроводности жидкости с потенциалом межчастичного взаимодействия Леннарда-Джонса). В

примере nemd [23] частицы находятся внутри параллелограмма, длина сторон которого не меняется, а углы изменяются (рис. 15). Изменение формы параллелепипеда происходит с помощью команды fix deform. Было проведено сравнительное испытание производительности примера LAMMPS nemd OpenMPI и LAMMPS nemd OpenTS DMPI. Входной файл был модифицирован, увеличены длины сторон параллелограмма.

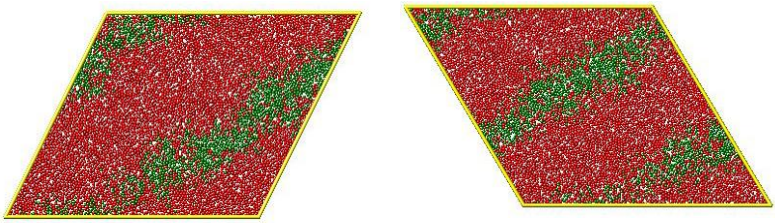


Рис. 15. Частицы находятся внутри параллелограмма, углы которого изменяются

20. Обтекание препятствий. В примере obstacle [23] в канале течения Пуазёйля находятся две пустоты, имеющие сферическую форму. Положение пустот не меняется со временем (рис. 16). Сферические пустоты в течении поддерживаются с помощью команды fix indent. Было проведено сравнительное испытание производительности примера LAMMPS obstacle OpenMPI и LAMMPS obstacle OpenTS DMPI. Входной файл был модифицирован, увеличен размер области симуляции и радиус пустот по сравнению с исходным примером.

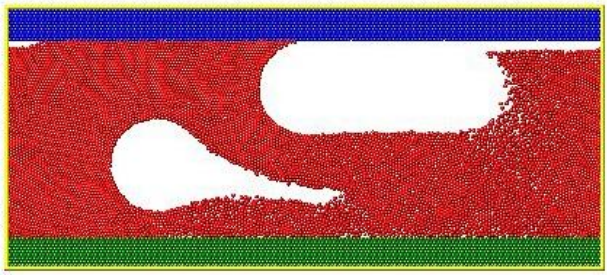


Рис. 167. Обтекание двух сферических препятствий

21. Перидинамика. Перидинамика – это нелокальное расширение механики сплошных сред, которая совместна с физической приро-

дой разрывов [52]. Основная идея перидинамики состоит в том, что локальное уравнение равновесия:

$$div\sigma + b = 0$$

заменяется уравнением:

$$\int_{\mathcal{H}_x} f(q,x)dV_q + b = 0,$$

где σ – поле напряжений, b – плотность объемной силы, f – плотность силы действия точки q на x, $\mathcal{H}_x$ – окрестность точки x радиуса δ. Если через y(x) обозначить деформацию, то:

$$\rho(x)\ddot{y}(x,t) = \int_{\mathcal{H}_x} f(q,x,t)dV_q + b(x,t),$$

где ρ– плотность. [52].

В примере peri [23] цилиндрическая мишень из упругого пластика поражается метательным снарядом (рис. 17). Для реализации модели используются команда atom_style peri и потенциал межчастичного взаимодействия peri. Обстрел цели симулируется с помощью команды fix indent. Результат обстрела подсчитывается с помощью команды compute damage. Эта команда реализована специально для модели перидинамики. Было проведено сравнительное испытание производительности примера LAMMPS peri OpenMPI и LAMMPS peri OpenTS DMPI.

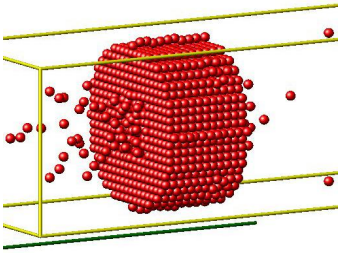


Рис. 17. В цилиндрическую мишень попал метательный снаряд

22. Дифракция электронов и рентгеновских лучей кристаллами Ni. Электронная и рентгеновская дифракция являются хорошо из-

вестными экспериментальными методами, используемыми для исследования химической структуры материала. В примере USER diffraction [23] реализован вычислительный метод облучения электронами и рентгеновскими лучами непосредственно с помощью атомистического моделирования без априорного знания элементарной ячейки. Этот метод применяется для изучения структуры (010) симметричных наклонных малоугловых и большеугловых границ зерен в Ni. Виртуальные электронные дифракционные картины и профили линий дифракции рентгеновских лучей показывают, что этот метод может различать малоугловые границы зерен с разными разориентациями и между малоугловыми границами с одинаковой разориентацией, но разными дислокационными конфигурациями. Для симметричного наклона границ зерен с совпадающими узлами Σ5 (210), Σ29 (520) и Σ5 (310) для решетки совпадающих узлов (010) виртуальные дифракционные методы могут выявить разориентацию границы зерен и показать незначительные различия между границами зерен [53]. Образец никеля виртуально облучался потоком электронов и рентгеновских лучей (рис. 18). Изображение получено с использованием программы VisIt [54] Ливерморской национальной лабораторией им. Э. Лоуренса.

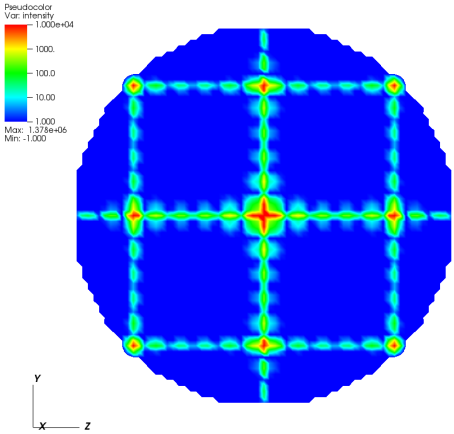


Рис. 18. Дифракция электронов и рентгеновских лучей кристаллами никеля

С помощью команды lattice fcc строится гранецентрированная кристаллическая решетка с длиной ребра 3.52 ангстрем. Масса каждого атома задается равной 5.71 г/моль. Создается кристалл в форме куба из 32000 атомов никеля. Потенциал межчастичного взаимодействия отсутствует. Симуляция облучения рентгеновскими лучами производится с

помощью команды compute xrd [55]. Симуляция облучения электронами производится с помощью команды compute SAED [56]. В источниках представлены соответствующие формулы и ссылки на дополнительные источники. Было проведено сравнительное испытание производительности примера LAMMPS USER diffraction OpenMPI и LAMMPS USER diffraction OpenTS DMPI.

23. Динамика диссипативных частиц. Динамика диссипативных частиц (DPD) – это метод мезомасштабных частиц, который устраняет разрыв между микроскопическим и макроскопическим моделированием. Его можно рассматривать как крупнозернистую модель молекулярной динамики, подходящую для больших масштабов времени и длины. DPD был успешно применен к различным областям приложений, особенно при моделировании гидродинамического поведения сложных жидкостей. Модель DPD разработана Хугербрюгге и Коэлменом [57]. Сила взаимодействия частиц i и j является суммой трех сил, консервативной, диссипативной и случайной:

$$F_{ij} = F_{ij}^{C} + F_{ij}^{D} + F_{ij}^{R}.$$

Одним из примеров, приведенных для модели DPD в LAMMPS, является пример, когда в системе с постоянным количеством частиц при постоянном объеме поддерживается постоянная энергия (ансамбль NVE). Соответствующие уравнения приведены в [58]. Координаты и вектора скоростей частиц считываются из входного файла. Было проведено сравнительное испытание производительности примера LAMMPS USER dpd OpenMPI и LAMMPS USER dpd OpenTS DMPI. Входной файл был модифицирован, увеличено количество временных шагов по сравнению с исходным примером.

24. Вычисление вязкости жидкости. В примере viscosity [23] вычисляется коэффициент вязкости жидкости с помощью формулы Ньютона:

$$\tau = -\eta \frac{\partial v}{\partial n},$$

где τ – вязкость, η – коэффициент вязкости, $\frac{\partial v}{\partial n}$ – градиент скорости вдоль оси, перпендикулярной к плоскости сдвига слоёв жидкости. В

процессе симуляции использовался сосуд с жидкостью в форме параллелограмма с изменяющимися углами и постоянными сторонами (рис. 15). Заданная температура поддерживается с помощью термостата Нозе-Хувера. Было проведено сравнительное испытание производительности примера LAMMPS viscosity OpenMPI и LAMMPS viscosity OpenTS DMPI. Входной файл был модифицирован, увеличена область симуляции.

25. Заключение. Технология создания гибридных программ, в которых сложная интеллектуальная часть реализуется при помощи динамически распараллеливаемого перебора на языке T++, а низкий уровень вычислений – при помощи традиционных низкоуровневых средств типа MPI или CUDA, обладают, по мнению авторов, большим потенциалом, который может быть востребован, в частности, при разработке интеллектуальных систем автоматизированного проектирования.

Проведенная проверка показала, что подход, выбранный для написания гибридного приложения, не ухудшает производительности реальных приложений. Разброс относительных изменений средних значений производительности составил от -4.9% до 5.8%, в среднем -0.2%, что пренебрежимо мало и приемлемо при тех преимуществах, которые получаются в результате переноса. Дальнейшие исследования будут посвящены разработке технологии адаптации пакетов с платформы OpenMPI на платформу OpenTS DMPI.

Литература

1. MPICH: A High-Performance, Portable Implementation of MPI. Argonne National Laboratory, Mathematics and Computer Science Division. URL: https://www.anl.gov/mcs/mpich-a-highperformance-portable-implementation-of-mpi.
2. MVAPICH: MPI over InfiniBand, Omni-Path, Ethernet/iWARP, and RoCE. The Ohio State University, Network-Based Computing Laboratory. URL: http://mvapich.cse.ohio-state.edu.
3. URL: www.cray.com.
4. TianHe-2A. URL: https://www.top500.org/system/177999.
5. Intel MPI. URL: https://software.intel.com/content/www/us/en/develop/tools/mpi-library.html.
6. Blue Gene/Q MPI. URL: http://www.redbooks.ibm.com/redbooks/pdfs/sg247948.pdf.
7. The IBM Parallel Environment (PE) Developer Edition. URL: http://www.redbooks.ibm.com/abstracts/tips1073.html
8. IBM Platform MPI. URL: https://www.ibm.com/support/knowledgecenter/en/SSENRW_4.2.0/get_started_admin/getting_started_mpi.html.
9. Installing SGI MPI packages. URL: https://downloads.linux.hpe.com/SDR/project/mpi/
10. Application Development Environment for Supercomputer Fugaku. URL: https://www.fujitsu.com/global/about/resources/publications/technicalreview/2020-03/article07.html

11. MS MPI. URL: https://docs.microsoft.com/en-us/message-passing-interface/microsoft-mpi.
12. MPI 4.0.URL: https://www.mpi-forum.org/mpi-40/
13. *Абрамов С.М., Васенин В.А., Мамчиц Е.Е., Роганов В.А., Слепухин А.Ф.*, Динамическое распараллеливание программ на базе параллельной редукции графов. Архитектура программного обеспечения новой версии Т-системы // Научная сессия МИФИ-2001, Сборник научных трудов. Т. 2, Москва, 22–26 января 2001 г., с. 234.
14. *Абрамов С.М., Кузнецов А.А., Роганов В.А*. Кроссплатформенная версия Т-системы с открытой архитектурой // Вычислительные методы и программирование, 8:1(2) (2007), с. 175–180, URL: http://num-meth.srcc.msu.ru/zhurnal/tom_2007/v8r203.html
15. *Кузнецов А.А., Роганов В.А.* Экспериментальная реализация отказоустойчивой версии системы OpenTS для платформы Windows CCS. // Труды Второй Международной научной конференции "Суперкомпьютерные системы и их применение (SSA'2008)" 27-29 октября 2008, г., Минск. — Минск: ОИПИ НАН Беларуси, 2008 с. 65-70 ISBN 978-985-6744-46-7
16. *Степанов Е.А.* Планирование в OpenTS — системе автоматического динамического распараллеливания. // М., МГИУ, сборник статей "Информационные технологии и программирование", выпуск 2, 2005.
17. *Абрамов С.М., Есин Г.И., Загоровский И.М., Матвеев Г.А., Роганов В.А.* Принципы организации отказоустойчивых параллельных вычислений для решения вычислительных задач и задач управления в Т-Системе с открытой архитектурой (OpenTS). // Международная конференция “Программные системы: теория и приложения (PSTA-2006)”, 23-28 октября 2006 г., г. Переславль-Залесский, Институт Программных Систем РАН, сборник трудов конференции, С. 257–264.
18. *Roganov V., Slepuhin A.* Distributed Extension of the Parallel Graph Reduction. GRACE: Compact and Efficient Dynamic Parallelization Technology for the Heterogeneous Computing Systems. International Conference on Parallel and Distributed Processing Techniques and Applications, June 25–28, 2001, Las Vegas, Nevada, USA.
19. *Moskovsky A., Roganov V., Abramov S.* Parallelism Granules Aggregation with the T-System. Parallel Computing Technologies: 9th International Conference, PaCT 2007 Pereslavl-Zalessky, Russia, September 2007. Proceedings. Victor Malyshkin (Ed.)- Berlin etc. Springer, 2007. – Lecture Notes in Computer Science: vol. 4671, pp. 293-302.
20. *Moskovsky A., Roganov V., Abramov S., Kuznetsov A.* Variable Reassignment in the T++ Parallel Programming Language. Parallel Computing Technologies: 9th International Conference, PaCT 2007 Pereslavl-Zalessky, Russia, September 2007. Proceedings. Victor Malyshkin (Ed.). Berlin etc. Springer, 2007. – Lecture Notes in Computer Science: vol. 4671, pp. 579-588.
21. LAMMPS. URL:https://lammps.sandia.gov.
22. Lennard-Jones, J. E. — Proc. Roy. Soc., 1924, v. A 106, p. 463.
23. LAMMPS example scripts. URL: https://lammps.sandia.gov/doc/Examples.html.
24. Gay J.G., Berne B.J. Modification of the overlap potential to mimic a linear site–site potential. Journal of Chemical Physics, 1981, vol. 74 pp. 3316-3319.
25. Потенциал Гея-Берне. URL:https://lammps.sandia.gov/doc/pair_gayberne.html
26. *Stuart S.J.; Tutein A.B.; Harrison J.A.* A reactive potential for hydrocarbons with intermolecular interactions. Journal of Chemical Physics, 2000, vol. 112, Issue 14, pp. 6472-6486.

27. Потенциал AIREBO. URL:https://lammps.sandia.gov/doc/pair_airebo.html.
28. *Brenner D.W., Shenderova O.A., Harrison J.A., Stuart S.J., Ni B., Sinnott S.B.* A second-generation reactive empirical bond order (REBO) potential energy expression for hydrocarbons. J Physics: Condensed Matter, 2002, vol. 14, 783-802.
29. *Hecht M., Harting J., Ihle T., Herrmann H.* Simulation of claylike colloids. Phys. Rev. E., 2005, vol. 72, 011408.
30. *Petersen M.K., Lechman J.B., Plimpton S.J., Grest G.S., Veld P.J., Schunk P.R.* Mesoscale Hydrodynamics via Stochastic Rotation Dynamics: Comparison with Lennard-Jones Fluid. J. Chem. Phys. 2010, vol. 132, 174106.
31. LAMMPS fix srd command. URL:https://lammps.sandia.gov/doc/fix_srd.html.
32. *Axilrod and Teller*, Interaction of the van der Waals type between three atoms. J. Chem. Phys., 1943, vol. 11, 299.
33. *Muto Y.* Nippon Sugaku. Buturigakkwaishi 17, 629 (1943).
34. *Бараш Ю.С., Гинзбург В.Л.* Некоторые вопросы теории сил Ван-дер-Ваальса. УФН, 1984, № 143 С. 345–389.
35. LAMMPS balance command. URL: https://lammps.sandia.gov/doc/balance.html.
36. LAMMPS fix balance command. URL: https://lammps.sandia.gov/doc/fix_balance.html.
37. *Shan T.R., Devine B.D., Kemper T.W., Sinnott S.B. Phillpot S.R.* Charge-optimized many-body potential for the hafnium/hafnium oxide system. Phys. Rev. B. 2010, vol. 81, 125328.
38. *Liang, T., Shan, T.R., Cheng, Y.T., Devine, B.D., Noordhoek M., Li Y., Lu Z., Phillpot S.R., Sinnott S.B.* Classical atomistic simulations of surfaces and heterogeneous interfaces with the charge-optimized many body (COMB) potentials. Materials Science and Engineering R: Reports, 2013, vol. 74(9), pp. 255-279.
39. *Horsfield P., Bratkovsky A.M., Fearn M., Pettifor D.G., Aoki M.* Bond-order potentials: Theory and implementation. Phys. Rev. B. 1996, vol. 53, 12694.
40. *Nose S.* A unified formulation of the constant temperature molecular-dynamics methods. Journal of Chemical Physics. vol. 81 (1), pp. 511–519.
41. *Hoover W.G.* Canonical dynamics: Equilibrium phase-space distributions. Phys. Rev. A. 31 (3): 1695–1697.
42. LAMMPS fix nvt command URL: https://lammps.sandia.gov/doc/fix_nh.html.
43. Mean squared displacement. URL: https://en.wikipedia.org/wiki/Mean_squared_displacement.
44. LAMMPS compute msd command. URL: https://lammps.sandia.gov/doc/compute_msd.html.
45. *Shinoda W., Shiga M., Mikami M.* Rapid estimation of elastic constants by molecular dynamics simulation under constant stress. Phys. Rev. B., 2004, vol. 69, 134103.
46. Calculate elastic constants. URL: https://lammps.sandia.gov/doc/Howto_elastic.html.
47. *Stillinger F.H., Weber T.A.* Computer simulation of local order in condensed phases of silicon, Phys. Rev. B, 1985, vol. 31, pp. 52-62.
48. *Peshl T., Ehvald P., Prandtl L.* Fizika uprugikh i zhidkikh tel. [Physics of elastic and fluid bodies] Moscow, Gostekhizdat, 1933. (In Russ.).
49. LAMMPS fix heat command. URL: https://lammps.sandia.gov/doc/fix_heat.html.
50. *Plimpton S.* Sandia National Labs, Modeling Thermal Transport and Viscosity with Molecular Dynamics. LAMMPS Users and Developers Workshop International Centre for Theoretical Physics (ICTP) March 2014 - Trieste, Italy. URL: https://www.lammps.org/tutorials/italy14/italy_kappa_viscosity_Mar14.pdf.
51. *Todd B., Daivis P.* Nonequilibrium Molecular Dynamics. Theory, Algorithms and Applications. Cambridge University Press, 2017.

52. *Silling S.A.* Peridynamics: Introduction. In: Voyiadjis G. (eds) Handbook of Nonlocal Continuum Mechanics for Materials and Structures. Springer, Cham, 2018
53. *Coleman S.P., Spearot D.E., Capolungo L.* Virtual diffraction analysis of Ni [010] symmetric tilt grain boundaries. Modelling and Simulation in Materials Science and Engineering, 2013, 21(5).
54. VisIt Open Source visualiztion software URL: https://wci.llnl.gov/simulation/computer-codes/visit/.
55. LAMMPS compute xrd command. URL: https://lammps.sandia.gov/doc/compute_xrd.html.
56. LAMMPS compute SAED command. URL: https://lammps.sandia.gov/doc/compute_saed.html.
57. *Hoogerbrugge, P.J; Koelman, J.M.V.* A Simulating Microscopic Hydrodynamic Phenomena with Dissipative Particle Dynamics. Europhysics Letters (EPL). 1992, 19 (3): 155–160.
58. *Larentzos J.P., Brennan J.K., Moore J.D., Mattson W.D.* LAMMPS Implementation of Constant Energy Dissipative Particle Dynamics (DPD-E), ARL-TR-6863, U.S. Army Research Laboratory, Aberdeen Proving Ground, MD, 2014.
59. *Everaers R., Ejtehadi M.R.* Interaction potentials for soft and hard ellipsoids. Phys. Rev. E, 2003, 67, 041710.
60. *Veld P., Plimpton S., Grest G.* Accurate and Efficient Methods for Modeling Colloidal Mixtures in an Explicit Solvent using Molecular Dynamics. Computer Physics Communications, 2008, 179(5).

Абрамов Сергей Михайлович — д-р физ.-мат. наук, чл.-корр. РАН, директор, Исследовательский центр мультипроцессорных систем, Институт программных систем имени А.К. Айламазяна РАН. Область научных интересов: Суперкомпьютерные технологии, сетевые технологии и IoT, метавычисления. Число научных публикаций — 174. abram@botik.ru; ул. Петра Первого, 4а, 152021, Ярославская обл., Переславский р-н, село Веськово, Россия; р.т.: +74852695228; факс: +74852695228.

Роганов Владимир Александрович — научный сотрудник, Исследовательский центр мультипроцессорных систем, Институт программных систем имени А.К. Айламазяна РАН. Область научных интересов: Суперкомпьютерные технологии, сетевые технологии. Число научных публикаций — 47. var@pereslavl.ru; ул. Петра Первого, 4а, 152021, Ярославская обл., Переславский р-н, село Веськово, Россия; р.т.: +74852695228; факс: +74852695228.

Осипов Валерий Иванович — канд. физ.-мат. наук, старший научный сотрудник, Исследовательский центр мультипроцессорных систем, Институт программных систем имени А.К. Айламазяна РАН. Область научных интересов: Суперкомпьютерные технологии, сетевые технологии. Число научных публикаций — 25. val@pereslavl.ru; ул. Петра Первого, 4а, 152021, Ярославская обл., Переславский р-н, село Веськово, Россия; р.т.: +74852695228; факс: +74852695228.

Матвеев Герман Анатольевич — научный сотрудник, Исследовательский центр мультипроцессорных систем, Институт программных систем имени А.К. Айламазяна РАН. Область научных интересов: Суперкомпьютерные технологии, сетевые технологии. Число научных публикаций — 22. gera@prime.botik.ru; ул. Петра Первого, 4а, 152021, Ярославская обл., Переславский р-н, село Веськово, Россия; р.т.: +74852695228; факс: +74852695228.

DOI 10.15622/ia.20.4.8

S. ABRAMOV, V. ROGANOV, V. OSIPOV, G. MATVEEV

IMPLEMENTATION OF THE LAMMPS PACKAGE ON THE T-SYSTEM WITH OPEN ARCHITECTURE

Abramov S., Roganov V., Osipov V., Matveev G. **Implementation of the LAMMPS Package on the T-System with Open Architecture.**

Abstract. Supercomputer applications are usually implemented in the C, C++, and Fortran programming languages using different versions of the Message Passing Interface library. The "T-system" project (OpenTS) studies the issues of automatic dynamic parallelization of programs. In practical terms, the implementation of applications in a mixed (hybrid) style is relevant, when one part of the application is written in the paradigm of automatic dynamic parallelization of programs and does not use any primitives of the MPI library, and the other part of it is written using the Message Passing Interface library. In this case, the library is used, which is a part of the T-system and is called DMPI (Dynamic Message Passing Interface). In this way, it is necessary to evaluate the effectiveness of the MPI implementation available in the T-system. The purpose of this work is to examine the effectiveness of DMPI implementation in the T-system. In a classic MPI application, 0% of the code is implemented using automatic dynamic parallelization of programs and 100% of the code is implemented in the form of a regular Message Passing Interface program. For comparative analysis, at the beginning the code is executed on the standard Message Passing Interface, for which it was originally written, and then it is executed using the DMPI library taken from the developed T-system. Comparing the effectiveness of the approaches, the performance losses and the prospects for using a hybrid programming style are evaluated. As a result of the conducted experimental studies for different types of computational problems, it was possible to make sure that the efficiency losses are negligible. This allowed to formulate the direction of further work on the T-system and the most promising options for building hybrid applications. Thus, this article presents the results of the comparative tests of LAMMPS application using OpenMPI and using OpenTS DMPI. The test results confirm the effectiveness of the DMPI implementation in the OpenTS parallel programming environment.

Keywords: dynamic parallelization, T-system with an open architecture, OpenTS, T++ programming language, molecular dynamics.

Abramov Sergey — Ph.D., Dr.Sci., Corresponding Member of RAS, Director, Research Center for Multiprocessor Systems, The Ailamazyan Program Systems Institute of the Russian Academy of Sciences. Research interests: Supercomputing technologies, network technologies and IoT, metacomputations. The number of publications — 174. abram@botik.ru; 4a, Petra Velikogo, 152021, Yaroslavl Region, Pereslavl area, Veskovo, Russia; office phone: +74852695228; fax: +74852695228.

Roganov Vladimir — Researcher, Research Center for Multiprocessor Systems, The Ailamazyan Program Systems Institute of the Russian Academy of Sciences. Research interests: Supercomputing technologies, network technologies. The number of publications — 47. var@pereslavl.ru; 4a, Petra Velikogo, 152021, Yaroslavl Region, Pereslavl area, Veskovo, Russia; office phone: +74852695228; fax: +74852695228.

Osipov Valeriy — Ph.D., Senior researcher, Research Center for Multiprocessor Systems, The Ailamazyan Program Systems Institute of the Russian Academy of Sciences. Research interests: Supercomputing technologies, network technologies. The number of publications — 25.

val@pereslavl.ru; 4a, Petra Velikogo, 152021, Yaroslavl Region, Pereslavl area, Veskovo, Russia; office phone: +74852695228; fax: +74852695228.

Matveev German — Researcher, Research Center for Multiprocessor Systems, The Ailamazyan Program Systems Institute of the Russian Academy of Sciences. Research interests: Supercomputing technologies, network technologies. The number of publications — 22. gera@prime.botik.ru; 4a, Petra Velikogo, 152021, Yaroslavl Region, Pereslavl area, Veskovo, Russia; office phone: +74852695228; fax: +74852695228.

References

1. MPICH: A High-Performance, Portable Implementation of MPI. Argonne National Laboratory, Mathematics and Computer Science Division. URL: https://www.anl.gov/mcs/mpich-a-highperformance-portable-implementation-of-mpi.
2. MVAPICH: MPI over InfiniBand, Omni-Path, Ethernet/iWARP, and RoCE. The Ohio State University, Network-Based Computing Laboratory. URL: http://mvapich.cse.ohio-state.edu.
3. URL: www.cray.com.
4. TianHe-2A. URL: https://www.top500.org/system/177999.
5. Intel MPI. URL: https://software.intel.com/content/www/us/en/develop/tools/mpi-library.html.
6. Blue Gene/Q MPI. URL: http://www.redbooks.ibm.com/redbooks/pdfs/sg247948.pdf.
7. The IBM Parallel Environment (PE) Developer Edition. URL: http://www.redbooks.ibm.com/abstracts/tips1073.html
8. IBM Platform MPI. URL: https://www.ibm.com/support/knowledgecenter/en/SSENRW_4.2.0/get_started_admin/getting_started_mpi.html.
9. Installing SGI MPI packages. URL: https://downloads.linux.hpe.com/SDR/project/mpi/
10. Application Development Environment for Supercomputer Fugaku. URL: https://www.fujitsu.com/global/about/resources/publications/technicalreview/2020-03/article07.html
11. MS MPI. URL: https://docs.microsoft.com/en-us/message-passing-interface/microsoft-mpi.
12. MPI 4.0.URL: https://www.mpi-forum.org/mpi-40/
13. Abramov S.M., Vasenin V.A., Mamchits E.E., Roganov V.A., Slepukhin A.F., [Dynamic parallelization of programs based on parallel graph reduction. Software architecture of the new version of the T-system] *Nauchnaya sessiya MIFI-2001, Sbornik nauchnykh trudov*. [Scientific session MEPhI-2001, Collection of scientific papers]. vol. 2, Moscow, 22–26 yan. 2001, p. 234. (In Russ.).
14. Abramov S.M., Kuznetsov A.A., Roganov V.A. [A cross-platform version of the T-system with an open architecture] *Vychislitel'nye metody i programmirovanie – Computational methods and programming*, 2007, vol. 8: 1(2), pp. 175-180. (In Russ.). URL: http://num-meth.srcc.msu.ru/zhurnal/tom_2007/v8r203.html.
15. Kuznetsov A.A., Roganov V.A. [Experimental implementation of a fault-tolerant version of the OpenTS system for the Windows CCS platform.] *Trudy Vtoroy Mezhdunarodnoy nauchnoy konferentsii "Superkomp'yuternye sistemy i ikh primenenie (SSA'2008)"* [Proceedings of the Second International Scientific Conference "Supercomputer Systems and Their Application (SSA'2008)"] October 27-29, 2008, Minsk. - Minsk: OIPI NAS Belarus, 2008 pp. 65-70 ISBN 978-985-6744-46-7. (In Russ.).
16. Stepanov E.A. [Scheduling in OpenTS, an automatic dynamic parallelization system] Moscow, MSIU, *sbornik statey "Informatsionnye tekhnologii i programmirovanie"* [Collection of articles "Information technologies and programming"], 2005, issue 2 (In Russ.).

17. Abramov S.M., Esin G.I., Zagorovskiy I.M., Matveev G.A., Roganov V.A. [The principles of organizing fault-tolerant parallel computing for solving computational and control problems in the T-System with an open architecture (OpenTS).] *Mezhdunarodnaya konferentsiya "Programmnye sistemy: teoriya i prilozheniya (PSTA-2006)",* [International conference "Software systems: theory and applications (PSTA-2006)"] October 23-28, 2006, Pereslavl-Zalessky, Program Systems Institute RAS, collection of conference proceedings, pp. 257–264. (In Russ.).
18. Roganov V., Slepuhin A. Distributed Extension of the Parallel Graph Reduction. GRACE: Compact and Efficient Dynamic Parallelization Technology for the Heterogeneous Computing Systems. International Conference on Parallel and Distributed Processing Techniques and Applications, June 25–28, 2001, Las Vegas, Nevada, USA.
19. Moskovsky A., Roganov V., Abramov S. Parallelism Granules Aggregation with the T-System. Parallel Computing Technologies: 9th International Conference, PaCT 2007 Pereslavl-Zalessky, Russia, September 2007. Proceedings. Victor Malyshkin (Ed.)- Berlin etc. Springer, 2007. – Lecture Notes in Computer Science: vol. 4671, pp. 293-302.
20. Moskovsky A., Roganov V., Abramov S., Kuznetsov A. Variable Reassignment in the T++ Parallel Programming Language. Parallel Computing Technologies: 9th International Conference, PaCT 2007 Pereslavl-Zalessky, Russia, September 2007. Proceedings. Victor Malyshkin (Ed.)- Berlin etc. Springer, 2007. – Lecture Notes in Computer Science: vol. 4671, pp. 579-588.
21. LAMMPS. URL:https://lammps.sandia.gov
22. Lennard-Jones, J. E. — Proc. Roy. Soc., 1924, v. A 106, p. 463.
23. LAMMPS example scripts. URL: https://lammps.sandia.gov/doc/Examples.html.
24. Gay J.G., Berne B.J. Modification of the overlap potential to mimic a linear site–site potential. *Journal of Chemical Physics,* 1981, vol. 74 pp. 3316-3319.
25. LAMMPS pair_style gayberne command URL: https://lammps.sandia.gov/doc/pair_gayberne.html
26. Stuart S.J.; Tutein A.B.; Harrison J.A. A reactive potential for hydrocarbons with intermolecular interactions. *Journal of Chemical Physics*, 2000, Vol. 112, Issue 14, pp. 6472-6486.
27. LAMMPS pair_style airebo command. URL: https://lammps.sandia.gov/doc/pair_airebo.html.
28. Brenner D.W., Shenderova O.A., Harrison J.A., Stuart S.J., Ni B., Sinnott S.B. A second-generation reactive empirical bond order (REBO) potential energy expression for hydrocarbons. *J Physics: Condensed Matter*, 2002, vol. 14, 783-802.
29. Hecht M., Harting J., Ihle T., Herrmann H. Simulation of claylike colloids. *Phys. Rev. E.*, 2005, vol. 72, 011408.
30. Petersen M.K., Lechman J.B., Plimpton S.J., Grest G.S., Veld P.J., Schunk P.R. Mesoscale Hydrodynamics via Stochastic Rotation Dynamics: Comparison with Lennard-Jones Fluid. *J. Chem. Phys.* 2010, vol. 132, 174106.
31. LAMMPS fix srd command. URL:https://lammps.sandia.gov/doc/fix_srd.html.
32. Axilrod and Teller, Interaction of the van der Waals type between three atoms. *J. Chem. Phys.*, 1943, vol. 11, 299.
33. Muto Y. Nippon Sugaku. Buturigakkwaishi 17, 629 (1943).
34. Barash Y.S., Ginzburg V.L. [Some questions of the theory of van der Waals forces] *UFN – Physics-Uspekhi,* 1984, no. 143 pp. 345–389 (In Russ.).
35. LAMMPS balance command. URL: https://lammps.sandia.gov/doc/balance.html.
36. LAMMPS fix balance command. URL: https://lammps.sandia.gov/doc/fix_balance.html.
37. Shan T.R., Devine B.D., Kemper T.W., Sinnott S.B. Phillpot S.R. Charge-optimized many-body potential for the hafnium/hafnium oxide system. *Phys. Rev. B.* 2010, vol. 81, 125328.
38. Liang, T., Shan, T.R., Cheng, Y.T., Devine, B.D., Noordhoek M., Li Y., Lu Z., Phillpot

S.R., Sinnott S.B. Classical atomistic simulations of surfaces and heterogeneous interfaces with the charge-optimized many body (COMB) potentials. *Materials Science and Engineering R: Reports*, 2013, vol. 74(9), pp. 255-279.
39. Horsfield P., Bratkovsky A.M., Fearn M., Pettifor D.G., Aoki M. Bond-order potentials: Theory and implementation. *Phys. Rev. B.* 1996, vol. 53, 12694.
40. Nose S. A unified formulation of the constant temperature molecular-dynamics methods. *Journal of Chemical Physics*. vol. 81 (1), pp. 511–519.
41. Hoover W.G. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A.* vol. 31(3), pp. 1695–1697.
42. LAMMPS fix nvt command URL: https://lammps.sandia.gov/doc/fix_nh.html.
43. Mean squared displacement. URL: https://en.wikipedia.org/wiki/Mean_squared_displacement.
44. LAMMPS compute msd command. URL: https://lammps.sandia.gov/doc/compute_msd.html.
45. Shinoda W., Shiga M., Mikami M. Rapid estimation of elastic constants by molecular dynamics simulation under constant stress. *Phys. Rev. B.*, 2004, vol. 69, 134103.
46. Calculate elastic constants. URL: https://lammps.sandia.gov/doc/Howto_elastic.html.
47. Stillinger F.H., Weber T.A. Computer simulation of local order in condensed phases of silicon, *Phys. Rev. B*, 1985, vol. 31, pp. 52-62.
48. Peshl T., Ehvald P., Prandtl L. *Fizika uprugikh i zhidkikh tel*. [Physics of elastic and fluid bodies] Moscow, Gostekhizdat, 1933. (In Russ.).
49. LAMMPS fix heat command. URL: https://lammps.sandia.gov/doc/fix_heat.html.
50. Plimpton S. Sandia National Labs, Modeling Thermal Transport and Viscosity with Molecular Dynamics. LAMMPS Users and Developers Workshop International Centre for Theoretical Physics (ICTP) March 2014 - Trieste, Italy. URL: https://www.lammps.org/tutorials/italy14/italy_kappa_viscosity_Mar14.pdf.
51. Todd B., Daivis P., *Nonequilibrium Molecular Dynamics. Theory, Algorithms and Applications.* Cambridge University Press, 2017.
52. Silling S.A. Peridynamics: Introduction. In: Voyiadjis G. (eds) Handbook of Nonlocal Continuum Mechanics for Materials and Structures. Springer, Cham, 2018
53. Coleman S.P., Spearot D.E., Capolungo L. Virtual diffraction analysis of Ni [010] symmetric tilt grain boundaries. Modelling and Simulation in Materials Science and Engineering, 2013, 21(5).
54. VisIt Open Source visualization software URL: https://wci.llnl.gov/simulation/computer-codes/visit/.
55. LAMMPS compute xrd command. URL: https://lammps.sandia.gov/doc/compute_xrd.html.
56. LAMMPS compute SAED command. URL: https://lammps.sandia.gov/doc/compute_saed.html.
57. Hoogerbrugge, P.J; Koelman, J.M.V. A Simulating Microscopic Hydrodynamic Phenomena with Dissipative Particle Dynamics. *Europhysics Letters (EPL).* 1992, 19 (3): 155–160.
58. Larentzos J.P., Brennan J.K., Moore J.D., and Mattson W.D. LAMMPS Implementation of Constant Energy Dissipative Particle Dynamics (DPD-E), ARL-TR-6863, U.S. Army Research Laboratory, Aberdeen Proving Ground, MD, 2014.
59. Everaers R., Ejtehadi M.R. Interaction potentials for soft and hard ellipsoids. *Phys. Rev*. E, 2003, 67, 041710.
60. Veld P., Plimpton S., Grest G. Accurate and Efficient Methods for Modeling Colloidal Mixtures in an Explicit Solvent using Molecular Dynamics. Computer *Physics Communications*, 2008, 179(5).

Руководство для авторов

Взаимодействие автора с редакцией осуществляется через личный кабинет на сайте журнала «Информатика и автоматизация» http://ia.spcras.ru/. При регистрации авторам рекомендуется заполнить все предложенные поля данных. Подготовка статьи ведется с помощью текстовых редакторов MS Word 2007 и выше или LaTex. Объем основного текста (до раздела Литература) - от 20 до 30 страниц включительно. Переносы разрешены. Номера страниц не проставляются. Основная часть текста статьи разбивается на разделы, среди которых являются обязательными: введение, хотя бы один «содержательный» раздел и заключение. Допускается также мотивированное содержанием и структурой материал а выделение подразделов. В основную часть опускается помещать рисунки, таблицы, листинги и формулы. Правила их оформления подробно рассмотрены на нашем сайте в разделе «Руководство для авторов».

Author guidelines

Interaction between each potential author and the Editorial board is realized through the pesoal account on the website of the journal “Informatics and Automation” http://ia.spcras.ru/. At the registration the authors are requested to fill out all data fields in the proposed form. The submissions should be prepared using MS Word 2007, LaTex. The text of the paper in the main part should not exceed 30 pages. Pages are not numbered; hyphenations are allowed. Certain figures, tables, listings and formulas are allowed in the main section, and their typography is considered in more detail at the journal web.

Signed to print 30.07.2021
Printed in Publishing center GUAP, 67, B. Morskaya, St. Petersburg, 190000, Russia

The journal is registered in the Russian Federal Agency for Communications and Mass-Media Supervision, certificate ПИ № ФС77-79228 dated September 25, 2020
Subscription Index П5513, Russian Post Catalog

Подписано к печати 30.07.2021. Формат 60×90 1/16. Усл. печ. л. 14,4. Заказ № 205.
Тираж 300 экз., цена свободная.
Отпечатано в Редакционно-издательском центре ГУАП, 190000, Санкт-Петербург, Б. Морская, д. 67

Журнал зарегистрирован Федеральной службой по надзору в сфере связи и массовых коммуникаций, свидетельство ПИ № ФС77-79228 от 25 сентября 2020 г.
Подписной индекс П5513 по каталогу «Почта России»