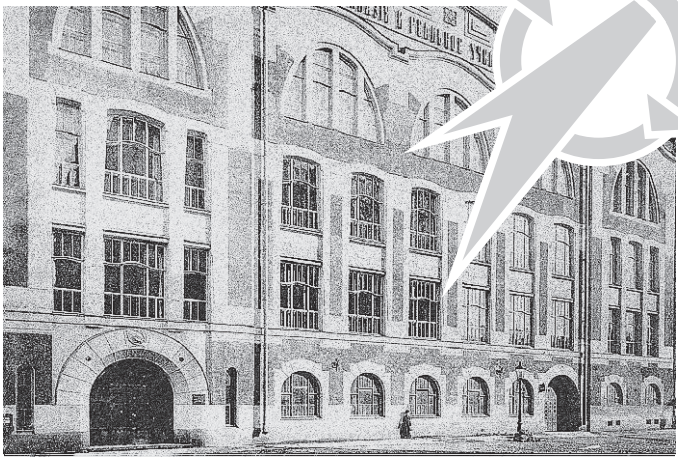
ISSN 2078-9181

РОССИЙСКАЯ АКАДЕМИЯ НАУК
Отделение нанотехнологий и информационных технологий

САНКТ-ПЕТЕРБУРГСКИЙ
ИНСТИТУТ ИНФОРМАТИКИ И АВТОМАТИЗАЦИИ РАН

ТРУДЫ СПИИРАН

proceedings.spiiras.nw.ru



ВЫПУСК 2(39)

Санкт Петербург
2015

18+

Труды СПИИРАН

Выпуск № 2(39), 2015

Научный, научно-образовательный, междисциплинарный журнал с базовой специализацией в области информатики, автоматизации и прикладной математики

Журнал основан в 2002 году

Учредитель и издатель

Федеральное государственное бюджетное учреждение науки
Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН)

Адрес редакции

199178, Санкт-Петербург, 14-я линия, д. 39,
e-mail: publ@iias.spb.su, сайт: http://www.proceedings.spiiras.nw.ru/

Подписано к печати 15.04.2015. Формат 60×90 1/16. Усл. печ. л. 15,4. Заказ № 105. Тираж 200 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 № ФС77-41695 от 19 августа 2010 г.
Подписной индекс 29393 по каталогу «Почта России»

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»

SPIIRAS Proceedings

Issue № 2(39), 2015

Scientific, educational, and interdisciplinary journal primarily specialized
in computer science, automation, and applied mathematics

Trudy SPIIRAN ♦ Founded in 2002 ♦ Труды СПИИРАН

Founder and Publisher

Federal State Budget Institution of Science
St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences
(SPIIRAS)

Editorial Board's address

14-th line VO, 39, SPIIRAS, St. Petersburg, 199178, Russia,
e-mail: publ@iias.spb.su, web: http://www.proceedings.spiiras.nw.ru/

Signed to print 15.04.2015
Printed in Publishing center GUAP, 67, B. Morskaya, St. Petersburg, 190000, Russia

The journal is registered in Russian Federal Agency for Communications and Mass-Media Supervision,
certificate ПИ № ФС77-41695 dated August 19, 2010 г.
Subscription Index 29393, Russian Post Catalog

СОДЕРЖАНИЕ

CONTENTS

УДК 004.8:007.3

Д.Н. Бирюков, А.Г. Ломако, Ю.Г. Ростовцев

ОБЛИК АНТИЦИПИРУЮЩИХ СИСТЕМ ПРЕДОТВРАЩЕНИЯ РИСКОВ РЕАЛИЗАЦИИ КИБЕРУГРОЗ

Бирюков Д.Н., Ломако А.Г., Ростовцев Ю.Г. **Облик антиципирующих систем предотвращения рисков реализации киберугроз.**
Аннотация. Предлагается облик интеллектуальных систем кибербезопасности со свойством антиципации. Обосновывается необходимость того, что система предотвращения компьютерных атак должна быть представлена в виде самообучающейся интеллектуальной системы самоорганизующихся гироматов. Ожидается, что применение искомых систем на практике, должно позволить более успешно решать задачи, связанные с предотвращением рисков реализации киберугроз.
Ключевые слова: антиципация, гиромат, киберсистема, предотвращение, семантика.

Biryukov D.N., Lomako A.G, Rostovtsev Y.G. **The Appearance of Anticipating Cyber Threats Risk Prevention Systems.**
Abstract. The appearance of intelligent cybersecurity systems featuring the property of anticipation is proposed. It is substantiated that the system of cyber attacks prevention should be designed in the form of self-learning intellectual self-organizing Gyromats system. It is expected that the application of the sought-for systems in practice would allow to solve problems related to prevention of cyber threats more efficiently.
Keywords: anticipation, gyromat, cybersystem, prevent, semantics

1. Введение. В настоящее время вопросам обеспечения информационной безопасности (ИБ) посвящается большое количество работ как у нас в стране, так и за рубежом. Но несмотря на это, ситуация складывается таким образом, что на сегодняшний день в арсенале средств обеспечения ИБ превалируют средства нейтрализации компьютерных атак (КА) на заключительных этапах их проявления, и практически отсутствуют механизмы, позволяющие осуществлять их упреждающее пресечение. Таким образом, на современном этапе развития средств обеспечения ИБ назрела объективная необходимость создания новых систем, способных осуществлять предотвращение возможных КА на защищаемые ресурсы [1]. Наиболее близкими по функционалу к искомым системам, являются системы предотвращения вторжений (СПВ). Однако, как показывает практика применения СПВ, они способны детектировать и частично отражать уже осуществляющиеся воздействия, но не в состоянии заблаговременно пресечь атакующие воздействия. Не способна к этому и вся совокупность существующих средств без соответствующей доработки и серьезного вмешательства квалифицированного специалиста в процесс их взаимного функционирования. Тем не менее вопрос, связанный с выработкой

решений, способствующих не вовлечению в рискованную ситуацию, или действий, предупреждающих вовлечение в нее, остается весьма актуальным (вопрос “предотвращения риска”). Следовательно актуальным является и вопрос создания системы, способной строить спецификации процессов упреждающего поведения в информационно-техническом конфликте.

2. Предотвращение – генеральная цель обеспечения ИБ. Можно утверждать, что цель взаимодействия экспертов с проектируемой системой в ходе предотвращения рисков должна сводиться к удовлетворению их информационных потребностей, связанных с получением от интеллектуальной системы (ИС), способной осуществлять порождение спецификаций упреждающего поведения в конфликте, знаний (сведений), необходимых для управления целями, задачами, рисками и проблемами в области обеспечения ИБ защищаемой организации. Чем большим количеством специальных знаний (адаптированных, содержащих модели возможных решений) обладает ИС, тем более полезна она может быть для экспертов в их деятельности, связанной с предотвращением риска.

Понятие «предотвращение» не является тривиальным. Как показывает анализ публикаций, не всегда понятие «предотвращение» истолковывается одинаково даже учеными единомышленниками. Ввиду этого видится необходимым декомпозировать рассматриваемое понятие на осмысленные составляющие.

Как видится, понятие «Предотвращение» (устранение ранними мерами) является агрегирующим понятием и основывается на понятиях «Обнаружение», «Предупреждение» и «Пресечение». При этом в рамках «Обнаружения» можно выделить «Узнавание» (распознавание) и «Открытие» (рассуждение с заключением), а в рамках «Предупреждения» – «Уведомление» (оповещение) и «Упреждение» (предварение).

Анализ возможностей средств обеспечения ИБ позволяет сделать вывод о том, что в настоящий момент наибольшее развитие получили средства, способные осуществлять распознавание известных атакующих воздействий (КА), оповещение должностных лиц о факте их совершения и пресечение КА. Значительно меньше развиты средства, способные осуществлять накопление и интеллектуальную обработку данных, приводящую к возможности порождения спецификаций упреждающего поведения.

Можно предположить, что способность системы обеспечения ИБ к упреждению в конфликте основывается на способности к манипулированию имеющимися у системы знаниями и способности к по-

рождению новых знаний (см. «Открытие»). Очевидно, что пополнять собственную базу знаний (БЗ) система обеспечения ИБ конкретной критической информационной инфраструктуры (КИИ) может либо информацией, которая предоставляется ей экспертами и программными (аппаратно-программными) средствами из ее состава, либо порожденными ею знаниями, тем самым сокращая время пополнения базы знаний, а следовательно и время выработки решений по обеспечению защищенности КИИ.

Пусть, например, к моменту времени t_0 система предотвращения КА накопила данные, необходимые и достаточные для построения моделей, возможно реализуемых в ходе информационно-технического конфликта (ИТК) процессов, схематично представленных в виде различных траекторий на рисунке 1. При этом различные процессы могут быть потенциально завершены с различными результатами. Некоторые результаты приемлемы для КИИ, а некоторые нет (см. оценки от «-3» до «+2»). Важным моментом является то, что система обеспечения ИБ КИИ может на отдельных этапах потенциально реализуемых процессов повлиять на их ход (см. узлы «B», «D», «E», «F», «I»). Очевидно, что чем быстрее система обеспечения ИБ КИИ смоделирует возможно реализуемые процессы, тем больше альтернатив упреждающего поведения она сможет предложить оператору.

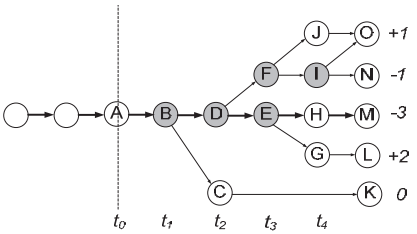


Рис. 1. Траектории возможно реализуемых в ходе ИТК процессов

Пусть наиболее вероятен процесс «A-M», который завершается с наихудшим исходом («-3»). Тогда, если системе обеспечения ИБ КИИ для моделирования процессов ИТК и выбора приемлемого варианта упреждающего поведения требуется $t_3 - t_0$ времени, то она не сможет предложить оператору ни одного приемлемого решения, если же системе потребуется менее чем $t_1 - t_0$ времени, то она сможет предложить целый ряд альтернатив. Таким образом, можно предположить, что чем более интеллектуальна система, тем в большем количестве и более скоро она способна выработать спецификации упреждающего поведения в ходе ИТК.

Следовательно, можно утверждать, что упреждающее поведение киберсистемы в конфликте сводится к синтезу такого сценария поведения, в ходе которого она способна изменить ход запланированного к реализации противником процесса, приводящего к негативным для КИИ последствиям. Для этого киберсистема должна быть способной изменить ход одного из мероприятий, являющегося составной частью процесса, который может быть завершен с неприемлемым для КИИ (киберсистемы) результатом. Изменение выполнения одного из мероприятий должно приводить к изменению процесса, а именно – к переходу к такой последовательности мероприятий (к траектории/трассе процесса), которая завершается допустимым для КИИ (киберсистемы) исходом.

Ввиду этого проблему исследования вопросов обеспечения защищенности КИИ в ходе ИТК предлагается свести к построению системы порождения сценариев упреждающего поведения в конфликте. Как видится, именно системы порождения сценариев упреждающего поведения в конфликте должны стать основой систем предотвращения компьютерных атак. При этом они не позиционируются как альтернатива существующим системам, обеспечивающим ИБ КИИ, а призваны лишь дополнить ее.

3. Антиципация – ключевой механизм упреждения в конфликте. На начальном этапе при поиске подходов к реализации в искомой системе упреждающего поведения предлагается обратить внимание на поведение биосистем, так как именно живые организмы и биосистемы выживают и эволюционируют уже много десятков миллионов лет, а поэтому видится целесообразным перенять часть опыта у них и заложить его в проектируемую систему обеспечения ИБ КИИ.

Проведенный анализ поведения живых существ и биосистем [2] позволил выявить ряд типовых механизмов их поведения, результатом которого является предотвращение возможных негативных для них последствий. Ряд механизмов упреждающего поведения изначально “вшит” в биоорганизмы. Это – рефлексы, иммунная система и другие. Однако наиболее целесообразному результативному упреждающему поведению животные и человек обучаются в ходе своей жизни, передавая и развивая его в поколениях. Очевидно, что одно из ведущих мест в процессе обучения в целом и в процессе разработки стратегий упреждающего поведения – в частности, играет головной мозг и механизмы мышления, реализованные в нем. Одним из примечательных механизмов является «*антиципация*». Так, например, Б.Ф.Ломов понимает антиципацию как «способность (в самом широком смысле) действовать и принимать те или иные решения с определенным вре-

менно-пространственным упреждением в отношении ожидаемых, будущих событий». Очевидно, что понятие антиципации может быть с пользой применимо не только к психической деятельности человека, но и к деятельности систем обеспечения информационной безопасности [3].

Можно предположить, что для того чтобы киберсистема обеспечения ИБ могла обладать свойством антиципации, она должна быть способной (см. рисунок 2): получать информацию через систему сенсоров (1.1, 1.2); оперировать информацией о прошлом опыте системы (2.1, 2.2); сопоставлять полученную информацию с имеющейся (3.1, 3.2); выдвигать гипотезы о возможных в перспективе событиях (4.1, 4.2); порождать стратегии целенаправленного поведения системы (5.1, 5.2); поддерживать требуемый уровень защищенности КИИ (6.1, 6.2).

Рис.2. Принципиальная схема антиципирующей системы

Интеллектуальную систему со свойством антиципации, способную осуществлять предотвращение атакующих воздействий на защищаемую КИИ, на начальном этапе практически невозможно описать полно и детально. Поэтому функционирование проектируемой системы предлагается рассмотреть через многомодельное представление процессов взаимодействия конфликтующих сторон и описать на нескольких стратах, приведенных в работе [1]. Понимание системы возрастает при последовательном переходе от одной страты к другой: чем ниже осуществляется спуск по иерархии, тем более детальным становится раскрытие системы, чем выше подъем, тем яснее становится смысл и значение всей системы.

Верхняя страта (искомая способность киберсистемы): предотвращение негативных компьютерных атак на защищаемую КИИ.

Нижняя страта (киберинтерпретация антиципации):

– сбор данных от сенсоров (рецепторов): распознавание элементарных явлений, регистрируемых специальными программными модулями, распределенными в сетевой инфраструктуре;

– классификация наблюдаемых элементарных явлений (агрегирование собираемых данных о наблюдаемых явлениях до уровня элементарных событий; формирование информационных признаков, присущих возможным опасностям);

– выявление типов потенциально возможных опасностей путем построения моделей потенциально реализуемых атакующей стороной процессов;

– определение существования задачи (задачи, требующей обращения на нее внимания) среди тех потенциальных опасностей, которые могут исходить от атакующей стороны;

– синтез схем потенциально возможного поведения системы предотвращения компьютерных атак (определение типовых решений, которые уже были выработаны системой предотвращения атакующих воздействий на более ранних этапах ее функционирования, либо были заложены в нее разработчиками; определение схем потенциально реализуемых вариантов решений, осуществляемое в случае отсутствия типовых решений);

– выбор конкретного варианта поведения из существующих (построенных): выбор наиболее подходящего варианта поведения для разрешения идентифицированного конфликта;

– построение стратегии реализации выбранного решения, направленного на упреждение потенциальных опасностей, на уровне операций;

– конструирование схем нейтрализации возможных угроз (конструирование схемы нейтрализации угроз на уровне операторов; формирование схемы управления системой сенсоров и эффекторов на уровне микроопераций, с целью пресечения компьютерных атак);

– верификация схемы управления эффекторами (и/или сенсорами).

Проведенный анализ показал, что как правило КИИ представляет собой разнородную (гетерогенную) распределенную сетевую инфраструктуру, а ее элементы потенциально уязвимы [3]. При этом уязвимости могут содержать различные как по назначению, так и по способу реализации элементы КИИ. Несмотря на это, проектируемая киберсистема предотвращения возможных атакующих воздействий (компьютерных атак) на КИИ ($\mathcal{K}$) должна быть потенциально способной функционировать в указанных условиях и учитывать дан-

ные факты при порождении сценариев упреждающего поведения в конфликте.

4. Самоорганизация – необходимое свойство системы, способной порождать сценарии упреждающего поведения в конфликте. Термин «самоорганизующаяся система» ввел У.Р. Эшби еще в середине прошлого века, но, к сожалению, до сих пор неизвестны достаточные условия, выполнение которых гарантировало бы начало самоорганизации. Сам же У.Эшби выделил два различных значения термина «самоорганизующаяся система» [4].

Во-первых, самоорганизация может заключаться в переходе «от системы с независимыми частями к системе с зависящими друг от друга частями» [4], при этом не учитывается, хороша или плоха возникающая организация. Системы такого рода Эшби предложил называть «самосвязующимися» (далее – «Самоорганизация_I»).

Во-вторых, самоорганизацией можно считать переход от плохой организации к хорошей, когда, например, ребенок, вначале потянувшись к огню, затем уже избегает его [4] (далее – Самоорганизация_II"). Правда, оговаривается Эшби: «…не существует «хорошей организации» в абсолютном смысле. Она всегда относительна...» [4].

Когда рассматривается Самоорганизация_II, то особенно важным является то, что самоорганизующаяся система сама переходит от "плохого" поведения к "хорошему".

Пусть множество состояний системы – S, а f – отображение S в S и определяется как множество пар (s_i, s_j), таких, что внутренняя движущая сила системы будет переводить состояние s_i в s_j. Если f является только функцией состояния (т.е. она может быть точно определена), то систему нельзя назвать "самоорганизующейся" [4].

Тогда, пусть функция f изменяема, а сами изменения не могут быть приписаны какой-либо причине во множестве S, то такой причиной может быть только некоторый внешний агент, воздействующий на систему S как ее вход. Если система должна быть в каком-то смысле "самоорганизующейся", понятие "само" должно быть расширено так, чтобы включать переменную α, причем для того, чтобы целое было ограничено, необходимо, чтобы причина изменений α находилась в S (или α). Таким образом, "самоорганизующейся" может быть только та система S, которая соединена с другой системой (из одной части). Тогда часть S может быть названа "самоорганизующейся" внутри целого $S+\alpha$. Только в этом частном и строго определенном

смысле можно признать, что система является “самоорганизующейся”, не будучи одновременно “самопротиворечивой” [4].

Утверждение 1. Необходимо, чтобы система $\mathcal{K}$ относилась к классу *самоорганизующихся систем*.

Доказательство:

“Самоорганизация_I”.

Лемма 1.1. Система $\mathcal{K}$ должна быть многоагентной

Доказательство.

КИИ представляет собой разнородную распределенную сетевую инфраструктуру [3], а ее элементы потенциально уязвимы, соответственно, необходимость наличия многоагентной системы предотвращения атакующих воздействий (КА) можно доказать следующим образом:

– пусть $OBG = \{obj_1, obj_2\}$ - множество взаимодействующих уязвимых элементов КИИ, причем obj_2 - целевой объект атаки;

– пусть $ACT = \{act_1, act_2, act_3\}$ - множество атакующих воздействий;

– $act_1(obj_1) = obj_1'$ - состояние элемента obj_1 КИИ после осуществления недопустимого воздействия act_1;

– пусть $act_2(act_1(obj_1), obj_2) = act_2(obj_1', obj_2) = obj_2'$ – двухэтапная атака;

– $act_2(obj_1, obj_2) = obj_2$ - выполнение второго этапа двухэтапной атаки при неуспешном выполнении первого этапа;

– $act_3(obj_1, obj_2) = act_3(obj_1', obj_2) = obj_2'$, где obj_2' - состояния элемента obj_2 КИИ после осуществления недопустимого воздействия act_3;

– пусть $IIPS = \{iips_a, iips_b, iips_c ...\}$ - элементы системы $\mathcal{K}$;

– пусть $P\_OBG = \{obj_1, obj_2, obj_3 ...\}$ - множество защищаемых ресурсов.

Допустим, система $\mathcal{K}$ может состоять из одного элемента:

– пусть $IIPS = \{iips_a\}$, т.е. $|IIPS| = 1$, а средство $iips_a$ - способно пресечь act_1;

– пусть $P\_OBG = \{obj_1\}$;

– тогда $iips_a\left(act_1\left(obj_1\right)\right) = obj_1$ означает, что $iips_a$ контролирует функционирование obj_1 и способно нейтрализовать атакующее воздействие $act_1\left(obj_1\right)$ в ходе его реализации. Очевидно, что и $act_2\left(iips_a\left(act_1\left(obj_1\right)\right), obj_2\right) = act_2\left(obj_1, obj_2\right) = obj_2$, однако act_3 способно воздействовать на obj_2 из состава КИИ (см. $OBG = \left\{obj_1, obj_2\right\}$ и $P\_OBG = \left\{obj_1\right\}$): $act_3\left(obj_1, obj_2\right) = obj_2'$, что является недопустимым. Следовательно, система предотвращения атакующих воздействий (КА) на КИИ в общем случае не может быть представлена системой, состоящей из одного элемента.

△ *Лемма 1.1.*

Лемма 1.2. Агенты системы $\mathcal{K}$ должны иметь возможность взаимодействовать друг с другом.

Доказательство.

Доказательство может быть выстроено на основе положений из двух ниже приведенных аксиом, сформулированных по результатам анализа порядка осуществления и пресечения атакующих воздействий.

Аксиома 1.2.1. Подавляющее большинство информационно-технических воздействий относится к классу многоэтапных.

Аксиома 1.2.2. Существуют мероприятия, связанные с предотвращением (пресечением) многоэтапных атакующих воздействий, при проведении которых результативность последующих этапов их выполнения зависит от результативности выполнения предыдущих этапов.

△ *Лемма 1.2.*

Простейший пример – киберсистема, состоящая из совокупности агентов сенсоров, измеряющих определенные параметры процессов, протекающих в киберпространстве, и агентов эффекторов, прерывающих определенные процессы из обнаруженных (на основании результатов измерений).

“Самоорганизация_II”.

Лемма 1.3. Система $\mathcal{K}$ для предотвращения новых видов воздействий должна быть способна порождать результативные стратегии поведения под влиянием внешней среды

Доказательство:

Для формулирования доказательства полезно ввести ряд обозначений, формализующих некоторые процессы, описывающие поря-

док функционирования киберсистемы с упреждающим поведением, реализуемым на основе антиципации:

– обнаружение воздействия ($act_i \in ACT$) производится через осуществление определенных измерений сенсорами системы: поступающие на вход(ы) системы в конкретный момент времени (T) данные (X), преобразовываются в данные (X_S), регистрируемые системой – $S: X \times T \to X_S$; S можно считать отображением, отвечающим за измерения;

– $Pr: D \times X_S \to L_{SM}$ – отображение регистрируемых данных (X_S) в формализованный вид представления (L_{SM}) для их последующего хранения и обработки (D задает порядок трансляции X_S в L_{SM});

– $M: L_{SM} \times Z_M \to Z_M$ – изменение (добавление) моделей (спецификаций, программ), которыми располагает система (изменение осуществляется под воздействием входных данных;

– $F: L_{SM} \times Z_M \to Z_{MeS}$, выбор стратегии поведения системы исходя из зарегистрированных данных и состояния системы (Z_{Mes} – модель действий системы);

– поскольку киберсистема должна быть способной не только принимать, но и передавать данные (а также осуществлять и другие активные действия), то необходимо реализовать отображение данных, хранящихся в системе в конкретный момент времени (T) и входящих в модель $Z_{Mes} \in Z_M$, в данные, формируемые на выходе системы (Y) – $R: D \times Z_{MeS} \times T \to Y$, где $D \times Z_{MeS} \to Y_S$.

Исходя из введенных выше обозначений, можно утверждать, что выбор стратегии поведения системой $\mathcal{K}$ в общем случае зависит от данных, поступающих из сторонней системы (систем) – см зависимость Z_{Mes} от X:

$$Z_{MeS} = F\left(L_{SM}, Z_M\right) = F\left(L_{SM}, M\left(L_{SM}, Z_M\right)\right) =$$
$$= F\left(Pr\left(D, X_S\right), M\left(Pr\left(D, X_S\right), Z_M\right)\right) =$$
$$= F\left(Pr\left(D, S\left(X, T\right)\right), M\left(Pr\left(D, S\left(X, T\right)\right), Z_M\right)\right).$$

△ *Лемма 1.3.*

Примечание: на формирование и выбор стратегии поведения системой предотвращения атакующих воздействий (КА) на КИИ также оказывают непосредственное влияние способности системы, связанные с преобразованием, накоплением и обработкой входных данных

во взаимосвязи с уже имеющейся информацией (с имеющимися моделями) – см. доказательство *Леммы 1.3*. Указанные способности реализуемы только в интеллектуальных системах.

▲ *Утверждение 1*.

5. Система порождения сценариев упреждающего поведения в конфликте – Интеллектуальная Система. Вопрос, связанный с тем, какую систему можно считать Интеллектуальной, остается открытым до сих пор. Ввиду этого, предлагается под Интеллектуальной Системой понимать такую систему, которая соответствует положениям, сформулированным В.К Финном [5, 6].

Интеллектуальные системы обладают специфической архитектурой, допускающей определенные вариации. Схематически эта архитектура может быть представлена следующим образом [5]:

ИС = (1) Решатель задач + (2) Информационная среда + (3) Интеллектуальный интерфейс.

(1) Решатель задач = (1.1) Рассуждатель + (1.2) Вычислитель + (1.3) Синтезатор.

(1.1) Рассуждатель реализует синтез и взаимодействие познавательных процедур, образующих автоматизированное рассуждение, областью применимости которого является класс задач, решаемых посредством формализованной эвристики. Логическим средством формализации этой эвристики и являются рассуждения. Правдоподобные рассуждения являются основным инструментом *Решателя* ИС реализуемым в *Рассуждателе* [6].

Существуют два типа *Рассуждателей*. *Рассуждатели первого типа* применимы к неизменяющемуся множеству исходных высказываний, характеризующих «замкнутый мир», а *Рассуждатели второго типа* реализуют формализованные эвристики для решения классов задач, исходными данными которых являются изменяемые и пополняемые множества высказываний (под изменением высказываний понимается пересмотр его истинностного значения, соответствующие базы фактов называют эпистемическими). *Рассуждатели второго типа* применимы к «открытым мирам», а их рассуждения называют когнитивными (правдоподобными) рассуждениями [5].

(1.2) Вычислитель применяется к числовым данным, используя численные методы, релевантные целям ИС.

(1.3) Синтезатор выбирает стратегии, адекватные не только цели ИС, но и состоянию БФ, и результатам предыдущих применений Решателя.

Если через Γ обозначить множество правил вывода, содержащих как правила достоверного вывода, так и правила правдоподобных

выводов, а через C – множество вычислительных процедур, то их комбинирование осуществляет *Синтезатор*.

(2) Информационная среда = (2.1) база фактов (БФ)+(2.2) база знаний (БЗ).

(2.1) БФ представляет рассматриваемую предметную область («замкнутый мир» или «открытый мир»; в первом случае БФ не изменяется, во втором – возможно ее пополнение в соответствии с результатами, полученными *Решателем* задач, и желаниями пользователя ИС как человеко-машинной системы).

Каждое элементарное событие – это элемент некоторого отношения. Фрагмент же предметной области характеризуется заданной системой отношений $R_1^{(k_1)},...,R_s^{(k_s)}$, с арностью $k_1,...,k_s$, соответственно.

Факт есть элементарное высказывание p_{ij} языка представления знаний L с некоторой оценкой v_{ij}, представляющее j-ый элемент отношения $R_i^{(k_i)}$, где $i = 1,...,s$. Таким образом, БФ есть множество элементарных высказываний p_{ij} с оценкой v_{ij}.

Наличие БФ как подсистемы ИС создает возможность осуществления машинного обучения [7], а, следовательно, расширения БЗ.

(2.2) *БЗ* – подсистема представления знаний.

Обычно выделяют три типа знаний для КС: декларативные, процедурные и концептуальные [6].

Под процедурными знаниями понимают задание алгоритмов и их комбинаций, применяемых в *Решателе* задач для достижения цели. Процедурным знанием являются стратегии решения задач, образованные посредством комбинирования различных видов, рассуждений и вычислений.

Под декларативным знанием понимают системы утверждений и, в частности, характеризацию предметной области (ПрО). Таковой являются аксиомы структуры данных (например, булевской) и дескриптивные утверждения, характеризующие предметную область (они могут быть необходимыми условиями корректности результатов применяемых процедур *Решателя* задач). Декларативным знанием ИС являются также утверждения, выражающие в импликативном виде правила вывода *Рассуждателя*. Эти утверждения образуют метатеорию ИС и создают возможность исследования на логическом уровне процедур *Рассуждателя*.

Концептуальным знанием ИС является множество утверждений и определений понятий, характеризующих принципы создания ИС.

Это знание является метатеоретическим, которым руководствуются создатели ИС.

(3) Интеллектуальный интерфейс включает в себя диалог (наилучший вариант – диалог на естественном языке), демонстрацию как результатов работы ИС, так и процесса их получения, графическое представление результатов, научение пользователя работе с ИС, поддержка интерактивного режима работы ИС.

Рассуждения и вычисления, представление знаний и интерфейс являются практическими реализациями принципов функционирования ИС. Посредством этих компонент функционирования ИС осуществляется интеллектуальная обработка данных.

Под «представлением знаний в ИС» понимают как выбор формы выражения знания посредством некоторого специального языка *L*, так и содержание, отображающего фрагмент предметной области, введенный в ИС в соответствии с целями, т.е. решаемыми задачами [8]. Наиболее известными формами представления знаний в ИС являются язык логики предикатов 1-го порядка, семантические сети и фреймы [7].

Утверждение 2. Самоорганизующаяся система $\mathcal{K}$ относится к классу Интеллектуальных Систем.

Доказательство:

Учитывая совокупность способностей, которыми должна обладать киберсистема, чтобы она могла быть отнесена к классу антиципирующих, а также детализацию процессов функционирования киберсистемы с упреждающим поведением [1], можно определить ряд параметров и отображений:

– $F_{ExtrW} : L_{SM} \times Z_M \to Z_{MeW}$, где Z_{MeW} – прогнозируемая модель наблюдаемого процесса (F_{ExtrW} – функция выявления типов потенциально возможных опасностей путем построения обобщенных моделей потенциально реализуемых атакующей стороной процессов);

– $Concl : Z_{MeW} \to C$, где C – оценка прогнозируемого процесса, наблюдаемого киберсистемой;

– $F_{ExtrS} : L_{SM} \times (Z_M \bigcup Z_{MeW}) \to Z_{MeS}$ – построение модели действий киберсистемы в условиях реализации прогнозируемого процесса, где Z_{MeS} – модель действий системы в ситуации Z_{MeW} (т.е. по результатам оценивания система должна быть способной принимать решения о том, какие ей необходимо осуществить действия для достижения потребного для нее и для защищаемой КИИ будущего; $F_{ExtrS} \subset F$);

– $F_{ExtrW} : L_{SM} \times (Z_M \cup Z_{MeS}) \to \bar{Z}_{MeW}$, где $\bar{Z}_{MeW}$ – прогнозируемая модель наблюдаемых процессов в условиях противодействия со стороны $\mathcal{K}$ (естественно предположить, что различные ответные действия интеллектуальной системы могут приводить к различным результатам, которые сама система должна заранее просчитывать);

– B – множество базовых элементов, через которые система сможет описывать предметную область конфликта;

– L – множество синтаксических правил построения более сложных структур, позволяющих описывать ПрО, используя B;

– $D = B \times L$, тогда $Pr : B \times L \times X_S \to L_{SM}$;

– Q – множество правил порождения киберсистемой новых знаний из имеющихся в конкретный момент (т.е. с учетом поступивших), $F_Q : Q \times Z_M \to Z_M$.

Лемма 2.1. Полнота и качество моделей упреждающего поведения в конфликте, формируемых системой $\mathcal{K}$, зависит от: имеющихся у нее знаний (Z_M), поступающих на ее вход(ы) данных (X), языка представления знаний (D) и правил порождения новых знаний из имеющихся (Q).

Доказательство:

$F_{ExtrW} : L_{SM} \times Z_M \to Z_{MeW}$ – потенциальная Задача (потенциально возможная опасность – см. [1]) – зависит от Z_M;

$F_{ExtrS} : L_{SM} \times (Z_M \cup Z_{MeW}) \to Z_{MeS}$ – потенциальное Решение (потенциально реализуемый вариант решения – см. [1]) – зависит от Z_M;

$F_Q : Q \times Z_M \to Z_M$, учитывая, что $M : L_{SM} \times Z_M \to Z_M$ и $Pr : B \times L \times X_S \to L_{SM}$, можно записать: $F_Q : Q \times B \times L \times X_S \times Z_M \to Z_M$, известно, что: $D = B \times L$, а $S : X \times T \to X_S$, тогда очевидно, что: полнота и качество моделей упреждающего поведения, зависящие от полноты и качества выявления потенциальных Задач (Z_{MeW}) и способов их потенциальных Решений (Z_{MeS}), зависят от Z_M, X, D и Q.

$\triangle$ *Лемма 2.1.*

В таблице 1 приведено соответствие элементов интеллектуальной системы и функций, которые должны быть реализованы в проектируемой киберсистеме.

Таблица 1. Соотнесение элементов ИС и функций $\mathcal{K}$

ИС	Система $\mathcal{K}$
1 Решатель задач	
1.1 Рассуждатель	$D = B \times L$; $F_Q : Q \times Z_M \to Z_M$; $F_Q : Q \times B \times L \times X_S \times Z_M \to Z_M$;
1.2 Вычислитель	Вычислитель – частный случай Рассуждателя при соответствующем Q
1.3 Синтезатор	$F_{ExtrW} : L_{SM} \times Z_M \to Z_{MeW}$; $F_{ExtrS} : L_{SM} \times (Z_M \cup Z_{MeW}) \to Z_{MeS}$; $Concl : Z_{MeW} \to C$; $F_{ExtrW} : L_{SM} \times (Z_M \cup Z_{MeS}) \to \overline{Z}_{MeW}$; $D \times Z_{MeS} \to Y_S$;
2 Информационная среда	
2.1 База Фактов	L_{SM} ;
2.2 База Знаний	$M : L_{SM} \times Z_M \to Z_M$; Z_M ;
3 Интеллектуальный интерфейс	$S : X \times T \to X_S$; $Pr : B \times L \times X_S \to L_{SM}$; $R : Y_S \times T \to Y$

▲ *Утверждение 2.*

Дополнительные требования к системе порождения сценариев упреждающего поведения в конфликте: (1) при реализации «*Рассуждателя*» в проектируемой системе необходимо учесть возможность его применения к «открытым мирам», а следовательно, (2) База Фактов, представляющая предметную область конфликтов, должна быть способной представлять “открытый мир”.

Выдвинутые требования (1, 2) обусловлены тем, что проектируемая система должна быть потенциально способной формировать сценарии упреждающего поведения в новых типах конфликтов (сценарии, предотвращения которых не вносились непосредственно в систему при ее создании) – см. Лемму 1.3.

Требование (3): «*Синтезатор*» должен выбирать стратегии поведения, адекватные не только состоянию информационной среды, но и целям системы, которые могут изменяться в ходе ее функционирования.

Если рассматривать различные цели как различные контексты, оказывающие влияние на принятие того или иного решения, то можно предположить, что контекст должен оказывать влияние на доступность решений. Пусть Br_b – доступность базовых элементов, а Br_l – проводимость связей между элементами B, тогда можно определить: $AS : Q \times B \times L \times X_S \times Z_M \times Br\_b \times Br\_l \to Z_M \times Br\_b \times Br\_l$.

А.Пуанкаре, Р.Курант, Г.Роббинс, Д.Пойа и И.Лакатос в своих работах выделяли, что в математическом творчестве важную роль играет аналогия. Можно предположить, что механизм порождения новых знаний (осуществляемых *Решателем* задач), основанный на выводах по аналогии (для обнаружения «Задач» и поиска их «Решений»), может быть весьма полезным для системы $\mathcal{K}$ в ходе порождения сценариев упреждающего поведения в конфликте.

Пусть:

– $F_{Z\_WW} : Z_M \times Z_{MeW} \to Z'_{MeW}$;

– $F_{Z\_SW} : Z_M \times Z_{MeS} \to Z'_{MeW}$;

– Z'_{MeW} – модели “Задач”, найденные по аналогии;

– $F_{Z\_WS} : Z_M \times Z_{MeW} \to Z'_{MeS}$;

– $F_{Z\_SS} : Z_M \times Z_{MeS} \to Z'_{MeS}$;

– Z'_{MeS} – модели “Решений”, найденные по аналогии.

Очевидно, что обнаружить «Задачи» и найти их «Решения», используя вывод новых знаний только по аналогии, можно не всегда, так как в ИС могут отсутствовать необходимые знания. Ввиду этого, видится полезным реализовать в «Синтезаторе» возможность порождения новых знаний путем комбинирования имеющихся. Очевидно, что произвольное комбинирование может привести к “комбинаторному взрыву” и к порождению моделей абсурдных процессов, поэтому в ИС должен быть реализован механизм направленного комбинирования для формирования моделей потенциально реализуемых процессов (т.е. $Q \subset Rl$, где Rl – правила вывода одних синтаксически и семантически верных конструкций, описывающих модели процессов, из других).

Проектированием ИС, способных корректировать собственные модели поведения под влиянием факторов из «Внешнего Мира», занимались и ранее в рамках исследований в области Искусственного Интеллекта. Подобные системы относятся к классу гироматов.

6. Киберсистема, способная к упреждающему поведению в конфликте, – самообучающаяся интеллектуальная система самоорганизующихся гироматов. Само слово «гиромат» придумано польским писателем-фантастом С. Лемом. По Лему, гиромат – это интеллектуальная машина, способная обнаруживать вокруг себя изменения и быстро откликаться на новизну, обучаться, меняя свое строение, приспосабливаясь к миру. Иными словами, гироматами С.Лем называет автоматы, самостоятельно составляющие для себя программу и «самоусовершенствующиеся». Далее идею гироматов, как устройств,

обладающих способностью изменять в соответствии с обстоятельствами свою семиотическую модель внешнего мира, научно обосновал и развил в своих работах Д.А. Поспелов [9, 10].

Гиромат Д.А. Поспелова – элементарная модель целесообразного поведения, способная адаптироваться к условиям решаемой задачи – уже содержал следующие «агенто-образующие» модули: блок мотивации; блок селекции (рецепторы); блок построения внутренней модели внешней среды; блок выдвижения гипотез; блок модельного опыта; блок выработки решений; блок активного опыта; блок времени.

Общая идея работы гиромата изложена в работах [9, 10]. Необходимыми условиями реализации искусственным агентом (гироматом) некоторого поведения являются наличие специальных устройств, непосредственно воспринимающих воздействия внешней среды (рецепторов) и исполнительных органов, воздействующих на среду (эффекторов), а также процессора (блока переработки информации) и памяти. Под памятью понимается способность агента хранить информацию о своем состоянии и состоянии среды. Таким образом, исходное представление о простейшем агенте Д.А.Поспелов свел к модели «организм-среда», описанной в монографии [9].

Поскольку проектируемая киберсистема ($\mathcal{K}$) должна быть интеллектуальной многоагентной системой, то возникает вопрос, связанный с исследованием процессов коммуникации, кооперации и координации агентов. При этом следует понимать, что распределенные интеллектуальные системы могут иметь единый орган управления, а в децентрализованных системах управление происходит только за счет локальных взаимодействий. В обоих этих случаях интеллектуальные процессы должны рассматриваться в контексте коллективного поведения, а центральным объектом исследования тогда становится группа или сообщество саморазвивающихся гироматов.

В работе [11] В.Б.Тарасов изложил заслуживающие внимания основы системно-организационного подхода в искусственном интеллекте (ИИ), включающие в себя следующие главные принципы:

– исследования интеллекта в иерархии взаимодействующих систем, что означает целесообразность изучения метаинтеллектуальных процедур, которые определяют, например, нормы взаимоотношений агентов в многоагентных системах;

– учета коллективной природы интеллекта, что предполагает обращение к семиотическим аспектам интеллекта;

– определения рекурсивных связей между интеллектом и деятельностью, согласно которому интеллект агента выступает как под-

система управления деятельностью, позволяющая ему организовать свои действия или действия другого агента;

– невозможности решения сложных задач отдельными агентами, опирающимися на локальные модели;

– дополнительности различных моделей интеллекта (аналогичный принципу Н.Бора), согласно которому невозможно отразить в одной модели многомерный характер понятия интеллект; для этого требуется построение семейства взаимодополняющих моделей;

– выделения системных единиц интеллекта.

В основу системного синтеза распределенного (децентрализованного) интеллекта целесообразно положить формирование функционально-структурной единицы как «универсального строительного блока» или «клетки» многоагентной системы. Системные единицы следует отличать от элементов: структурный элемент - это простейшая, неделимая часть системы, которая обычно не сохраняет свойства системы как целого, тогда как важнейшим требованием к функционально-структурной единице является сохранение важнейших свойств организации всей системы.

Исходя из результатов вышеприведенных рассуждений, будем полагать, что искомая интеллектуальная система, способная к порождению спецификаций упреждающего поведения в конфликте, может быть представлена в виде иерархии взаимодействующих частично-упорядоченных гироматов.

Следует отметить, что иерархия взаимодействующих гироматов тоже есть гиромат, но обладающий более совершенными «агентообразующими» модулями по сравнению с отдельно взятыми гироматами, входящими в иерархию.

В целом же, проектируемая киберсистема должна быть в состоянии как обнаруживать потенциально опасные процессы – «Задачи», так и находить пригодные их “Решения”. Очевидно, что обладая пустой Базой Знаний, система будет не в состоянии решить поставленные задачи. Поэтому видится необходимым указать, что система для решения перечисленных выше задач должна обладать необходимым объемом исходных знаний. Точно определить необходимый и достаточный объем знаний априорно невозможно, так как априорно неизвестны конкретные задачи, которые могут возникнуть перед киберсистемой в ходе защиты КИИ.

7. Заключение. Можно предположить, что информационно-технические системы, обладающие свойством антиципации и способные синтезировать спецификации упреждающего поведения в конфликте, в скором будущем найдут широкое применение в области

обеспечения безопасности компьютерных систем, входящих в состав КИИ, а также и в других областях деятельности человека.

Очевидно, что синтез и применение антиципирующих систем упреждения атакующих воздействий (компьютерных атак), должны повысить уровень защищенности критической информационной инфраструктуры. Сами искомые системы должны быть реализованы в виде многоагентных интеллектуальных самоорганизующихся систем, которые могут быть представлены в виде иерархии взаимодействующих гироматов. Как видится, именно гироматы должны стать основой антиципирующих систем предотвращения рисков реализации киберугроз.

Литература

1. *Бирюков Д.Н., Ломако А.Г.* Подход к построению системы предотвращения киберугроз // Проблемы информационной безопасности. Компьютерные системы. 2013. №2. С. 13–19.
2. *Бирюков Д.Н.* Анализ способностей живых организмов при проектировании систем кибербезопасности // Методы обеспечения информационной кибербезопасности. Труды ИСА РАН. М.: КомКнига. 2013. Т. 27 (доп. выпуск). С. 431–446.
3. *Бирюков Д.Н., Ломако А.Г.* Построение систем информационной безопасности: от живых организмов к киберсистемам // Защита информации. INSIDE. 2013. №2. С. 2–6.
4. *Эшби У.Р.* Принципы самоорганизации // Принципы самоорганизации // М.: Мир. 1966. С. 314–343.
5. *Финн В.К.* Об интеллектуальном анализе данных // Новости Искусственного интеллекта. 2004. № 3. С. 3–18.
6. *Финн В.К.* Искусственный интеллект: Идейная база и основной продукт // Труды 9-ой национальной конференции по искусственному интеллекту. М.: Физматлит. 2004. Т. 1. С. 11–20.
7. *Jain S.* Systems That Learn // An Introduction to Learning Theory, second edition. The MIT Press. Cambridge. Massachusetts. London. England. 1999.
8. *Nilsson N.J.* Artificial Intelligence: A New Synthesis // Morgan Kaufmann Publishers. Inc. San Francisco. California. 1998. 513 p.
9. *Гаазе-Рапопорт М.Г.* От амебы до робота: модели поведения // М.: Наука. 1987. 286 с.
10. *Поспелов Д.А.* Мышление и автоматы // М.: Советское радио. 1972. 224 с.
11. *Тарасов В.Б.* Системно-организационный подход в искусственном интеллекте // Программные продукты и системы. 1999. №3. С. 6–13.

References

1. Biryukov D.N., Lomako A.G. [Approach to creation of system of cyber-threats preventing]. *Problemy informatsionnoy bezopasnosty. Kompyuternie sistemy – Problems of information security. Computer systems*. 2013. no. 2. pp. 13–19. (In Russ).
2. Biryukov D.N. [Analysis of the ability of living organisms in the design of systems cybersecurity]. *Metody obespeceniya informatsionnoy kiberbezopasnosty. Trudy ISA RAN – ISA RAS proceedings Methods of providing information cybersecurity.* M.: KomKniga. 2013. vol. 27 (add. issue). pp. 431–446. (In Russ).
3. Biryukov D.N., Lomako A.G. [Design and construction of information security from living organisms to cybersystems]. *Zashita informatiyi – Data protection. INSIDE*. 2013. no. 2. pp. 2–6. (In Russ).
4. Ashby W.R. [Principles of self-organization]. *Principy samoorganizacii – Principles of self-organization.* M.:Mir. 1966. pp. 314–343.

5. Finn V.K. [About data mining]. *Novosty isskustvennogo intellekta – Artificial intelligence news*. 2004. no. 3. pp. 3–18. (In Russ).
6. Finn V.K. [Artificial intelligence: a Conceptual framework and the main product]. *Trudy 9-oj nacional'noj konferencii po iskusstvennomu intellektu* [Proceedings of the 9th national conference on artificial intelligence]. M.: Fizmatlit. 2004. vol. 1. pp. 11–20. (In Russ).
7. Jain S. Systems That Learn. An Introduction to Learning Theory, second edition. The MIT Press. Cambridge. Massachusetts. London. England. 1999.
8. Nilsson N.J. Artificial Intelligence: A New Synthesis. Morgan Kaufmann Publishers. Inc. San Francisco, California. 1998. 513 p.
9. Gaaze-Rapoport M.G. *Ot ameby do robota: modeli povedenija* [From the amoeba to the robot: model behavior]. M.: Nauka. 1987. 286 p. (In Russ).
10. Pospelov D.A. *Myshlenie i avtomaty* [Thinking and machines]. M.: Sovetskoe radio. 1972. 224 p. (In Russ).
11. Tarasov V.B. [Systematic organizational approach in artificial intelligence]. *Programmniyi produkty i sistemy – Software and systems.* 1999. no. 3. pp. 6–13.

Бирюков Денис Николаевич — к-т техн. наук, профессор кафедры систем сбора и обработки информации, Военно-космическая академия имени А.Ф. Можайского. Область научных интересов: системный анализ, защита информации, интеллектуальная поддержка принятия решений. Число научных публикаций — 70. Biryukov.D.N@yandex.ru; ул. Ждановская, д. 13, г. Санкт-Петербург, 197198; р.т.: (812) 237-19-60.

Biryukov Denis Nikolaevich — Ph.D., professor of systems for collecting and processing information department, Mozhaisky Military Space Academy. Research interests: system analyses, IT-Security, intelligent decision support. The number of publications — 70. Biryukov.D.N@yandex.ru; 13, Zhdanovskaya street, St.-Petersburg, 197198, Russia; office phone: (812) 237-19-60.

Ломако Александр Григорьевич — д-р техн. наук, профессор кафедры систем сбора и обработки информации, Военно-космическая академия имени А.Ф. Можайского. Область научных интересов: информационная безопасность, теоретическое и системное программирование, синтез и верификация корректности моделей программ. Число научных публикаций — 250. lomako_ag@mail.ru; ул. Ждановская 13, 197198, Санкт-Петербург; р.т.: +7(812) 237-19-60.

Lomako Aleksandr Grigor'evich — Ph.D., Dr. Sci., professor of system for collecting and processing information department, Mozhaisky Military Space Academy. Research interests: information security, theoretical and system programming, synthesis and verification of program models. The number of publications — 250. lomako_ag@mail.ru; 13, Zhdanovskaya street, St.-Petersburg, 197198, Russia; office phone: +7(812) 237-19-60.

Ростовцев Юрий Григорьевич — д-р техн. наук, профессор, заслуженный деятель науки и техники Российской Федерации, заслуженный работник высшей школы Российской Федерации, профессор кафедры систем сбора и обработки информации, Военно-космическая академия имени А.Ф. Можайского. Область научных интересов: системный анализ, теоритическая и прикладная кибернетика, методология знакового моделирования, радиотехника. Число научных публикаций — 350. Y.Rostovtsev@yandex.ru; ул. Ждановская 13, Санкт-Петербург, 197198; р.т.: +7(812) 237-19-60.

Rostovtsev Yuriy Grigorievich — Ph.D., Dr. Sci., professor, honored scientist and technology of Russian Federation, honored worker of higher school of Russian Federation, professor of systems for collecting and processing information department, Mozhaisky Military Space Academy. Research interests: system analyses, theoretical and applied cybernetics, the methodology of symbolic modeling, radio engineering. The number of publications — 350. Y.Rostovtsev@yandex.ru; 13, Zhdanovskaya street, St.-Petersburg, 197198, Russia; office phone: +7(812) 237-19-60.

РЕФЕРАТ

Бирюков Д.Н., Ломако А.Г., Ростовцев Ю.Г. **Облик антиципирующих систем предотвращения рисков реализации киберугроз.**

На современном этапе развития средств обеспечения информационной безопасности назрела объективная необходимость создания систем, способных осуществлять предупреждение и заблаговременное пресечение компьютерных атак на защищаемые ресурсы. Так же можно наблюдать, что в области безопасности компьютерных систем и сетей с каждым днем все чаще упоминаются и рекламируются различные биоинспирированные подходы, основанные на биологической метафоре. В развитии биоинспирированных подходов предлагается наделить комплексные средства обеспечения информационной безопасности принципиально новым свойством, позволяющим им предвидеть развитие событий, явлений, результатов действий и готовиться к ним. Такое свойство называется “Антиципация”.

В работе обоснованы основные задачи, решение которых должно позволить киберсистеме осуществлять предотвращение компьютерных атак. Определено, что система, способная предотвращать компьютерные атаки, должна относиться к классу интеллектуальных самоорганизующихся систем и быть представлена в виде иерархии взаимодействующих гироматов.

SUMMARY

Biryukov D.N., Lomako A.G., Rostovtsev Y.G. **The Appearance of Anticipating Cyber Threats Risk Prevention Systems.**

At the present stage of development of information security there is objective necessity of developing systems capable of carrying out prevention and early prevention of cyber attacks on protected resources. You can also observe that the security of computer systems and networks with each passing day more and more often mentioned and advertised bioinspired different approaches based on biological metaphor. In the development of bioinspired approach the means of the complex information security to endow are offered by fundamentally new feature that allows them to anticipate developments, events, results of operations and prepare for them. This property is called "Anticipation".

In work the main objectives, which decision has to allow cybersystem to carry out prevention of computer attacks, are proved. It is defined that the system capable to prevent computer attacks has to belong to the class of the intellectual self-organizing systems and to be presented in the form of hierarchy of the interacting gyromats.

УДК 004.056.53

С.А. Петренко

МОДЕЛЬ КИБЕРУГРОЗ ПО АНАЛИТИКЕ ИННОВАЦИЙ DARPA

Петренко С.А. **Модель киберугроз по аналитике инноваций DARPA.**
Аннотация. В работе рассматривается задача определения актуальной модели киберугроз цифровой обработки данных по аналитике инноваций DARPA. С 2002 года агенство DARPA проводит широкий спектр научных исследований для достижения и сохранения технологического превосходства вооруженных сил США в киберпространстве. В соответствии с этим задача определения актуальной модели киберугроз цифровой обработки данных рассматривается как адаптивная коррекция современных киберугроз по текущим НИОКР DARPA.
Ключевые слова: научные исследования, объект информатизации, киберугрозы, модель киберугроз, инновации, кибербезопасность..

Petrenko S.A. **Model Cyber Threats by Analysis of DARPA Innovations.**
Abstract. This paper considers the problem of determining the actual model cyber digital data analytics innovation DARPA. Since 2002 DARPA agency conducts a wide range of research to achieve and maintain the technological superiority of the US military in cyberspace. According to this research, the problem of determining the actual cyber digital data processing is regarded as adaptive correction of modern cyber threats on current R&D DARPA.
Keywords: research, cybersecurity, cyber threat, cyber model, innovation, the object information.

1. Введение. В настоящее время ряд технологически развитых государств (более 20 стран) продекларировали разработку «кибероружия». В США в декабре 2011 года от Конгресса было получено разрешение на развитие «наступательного» кибероружия. Во Франции в 2008 году в «Белой книге по обороне и национальной безопасности» введено понятие «кибервойна» и раскрыты ее составляющие – «кибероборона» и «наступательные возможности для кибервойны». В Германии в феврале 2011 года принята «Стратегия безопасности в киберпространстве». Аналогичный документ введен в действие в Великобритании с ноября 2011 года.

Для разработки «кибероружия» упомянутыми странами проводится широкий спектр НИОКР. В частности, в США ведутся работы по созданию специальных программно-аппаратных комплексов: *«PRISM» (сбор и обработка метаданных), «Feed through», «Gourmet through» и «Jet p low» (дистанционное внедрение «закладок» в персональные компьютеры), «Quantum Insert » (перенаправление трафика к ложным сайтам Интернета), «Dropout Jeep» (дистанционный съем информации с айфонов фирмы «Apple»), «Monkey calendar» (sms-сообщения о местонахождении мобильных телефонов), «Rage master» (перехват информации с экранов компьютеров), «Genie»*

(контроль функционирования 85000 «закладок-шпионов» по всему миру). На примере инноваций DARPA покажем возможные современные направления поисковых исследований в области кибербезопасности и определим модель новых киберугроз.

2. Проекты DARPA. Катастрофические последствия 11 сентября продемонстрировали неспособность США к отражению подобных терактов, беспрецедентных как по своему масштабу, так и асимметричности вызовов безопасности. Для разрешения сложившейся ситуации DARPA инициировала ряд специальных НИОКР [1–3] с бюджетом финансирования от 500 млн (2002-2005 годы) до 1 млрд долл. (2002-2009 годы). Проект «Предупреждение террористических актов» (Terrorism Information Awareness - TIA) позволил на основе анализа большого количества разнородных данных о слабо связанных между собой событий, таких как покупка авиа- и железнодорожных билетов, бронирование номеров в гостиницах, покупка химикатов и взрывчатых веществ, приобретение огнестрельного оружия и др., выявлять преступные группы лиц, готовящихся совершить террористический акт с применением оружия массового уничтожения (ядерного, химического, биологического) на территории США. Для реализации проекта были привлечены модели и методы системного анализа, исследования операций, теории игр, теории вероятности и статистического анализа, теории принятия решений и пр. Другой проект DARPA был направлен на разработку специализированного программного обеспечения так называемого "ситуационного анализа" (Software for Situational Analysis), который позволил в автоматизированном режиме: распознавать людей на расстоянии, обнаруживать противника, осуществляющего наблюдение за целями (объектами критической инфраструктуры) на территории США; автоматически находить, извлекать и связывать между собой отрывочные и фрагментарные представления о намерениях и деятельности групп людей, содержащиеся в больших массивах открытых и закрытых источников информации; достаточно точно моделировать субъективные представления и социальное поведение малочисленных по составу групп для имитации и проигрывания асимметричных действий противника; обеспечивать более эффективные средства анализа и принятия решения для пресечения преступной деятельности. Проект «Моделирование асимметричных воздействий» (Wargaming the Asymmetric Environment - WAE) позволил выявлять мотивы и своевременно раскрывать замысел террористических действий. В результате были созданы имитационные модели поведения отдельных людей и небольших групп с учетом их психологии, культуры, политических взглядов, уровня образования и жизненного

опыта (Scalable Social Network Analysis - SSNA). Также были разработаны имитационные модели поведения отдельных враждебно настроенных стран, их ключевых политических лидеров и террористических групп. Кроме того, построены аналитические модели для принятия решений, позволяющие прогнозировать различные ситуации в реальном масштабе времени (Rapid Analytical War Gaming - RAW). Здесь был применен математический аппарат теории игр со смешанными стратегиями, а также теория принятия решений в условиях неопределенности. Для повышения эффективности и координации совместных действий американских спецслужб по своевременному обнаружению террористов, раскрытию их замыслов и предотвращению терактов DARPA инициировала проекты «Генуя» и «Генуя-2». В результате была создана так называемая "динамическая виртуальная среда" для снятия возможных организационных и технических барьеров в совместной работе специалистов различных ведомств и организаций. В основу были положены модели и методы нечеткого структурирования аргументов, трехмерной цветной визуализации и организации адаптивной памяти.

Для обеспечения устойчивости и живучести критически важных объектов государственного и военного управления в чрезвычайных условиях DARPA инициировала долгосрочную программу – «Научные и инженерные методы» (Information Assurance Science and Engineering Tools - IASET), - которая объединила усилия специалистов в смежных областях знаний (исследование операций, системотехника, вычислительные системы и сети, кибербезопасность, операционные системы, базы данных и др.).

Проект «Безопасные и живучие информационные системы» (Organically Assured and Survivable Information Systems - OASIS) позволил выработать новые архитектурные решения комплексной системы защиты критических важных информационных систем. В результате была создана новая клиент-серверная технология обеспечения устойчивости и живучести вычислительных систем на основе современных методов обнаружения вторжений, адаптивной защиты, отказоустойчивости и реконфигурации.

Для оперативного контроля и прогнозирования состояния критически важных информационных систем реализован проект «Новые методы обнаружения кибератак» (Advanced Network Surveillance), в рамках которого созданы новые технологии обнаружения массовых и групповых кибератак. Создан прототип самообучающейся системы контроля и прогнозирования состояния критически важных информационных систем в условиях воздействия противника.

В рамках проекта «Корреляционный анализ кибернападения» (Cyber Attack Data Correlation) были разработаны адаптивные методы корреляционной обработки и классификации регистрируемых данных о состоянии критически важных информационных систем в условиях массовых враждебных программно-математических воздействий. Основное назначение – использование в крупных территориально-распределенных вычислительных сетях для определения фактов скоординированного широкомасштабного кибернападения и последующего адекватного противодействия.

В 2014 году на сайте DARPA (www.darpa.mil/Our_Work/I2O/Programs) был опубликован актуальный перечень поисковых программ исследований [3], в том числе:

- Профилирование поведения пользователей (Active Authentication);
- Активная киберзащита с ложными целями (ACD);
- Обнаружение аномальных процессов в обществе (ADAMS);
- Автоматизированный анализ кибербезопасности (APAC);
- Инфракрасный страж (ARGUS-IR);
- Высоконадежный семантический транслятор (BOLT);
- Адаптивная система (CRASH);
- Поисковая компьютерная система (CSSG);
- Формальная верификация (CSFV);
- Противодействие инсайдерам (CINDER);
- Глубокая очистка контента (DEFT);
- Психофизическая защита (DCAPS);
- Высоконадежная киберзащита беспилотных летательных аппаратов (HACMS);
- Семантический анализ (ICAS);
- Поиск контента (Memex);
- Повышение устойчивости программного обеспечения (MUSE);
- Транспарентные вычисления (Transparent Computing);
- Создание «живучего облака» (MRC);
- Рубрикация и семантическая классификация документов (MADCAT);
- Боевые операции в киберпространстве (Plan X);
- Надежное программирование (PPAML);
- Криптозащита вычислений ((PROCEED);
- Автоматизированный перевод речи (RATS);
- Безопасные коммуникации (SAFER);

- Работа в социальных СМИ (SMISC);
- Программа контроля НДВ в закупаемом для нужд Минобороны США ПО (VET);
- Наглядная визуализация (VMR);
- Распознавание сетей (WAND);
- Большие данные (XDATA) и пр.

Например, *программа MUSE* предназначена для повышения надежности и безопасности прикладного программного обеспечения на основе методов надежности программ, машинного обучения и формальной верификации программного обеспечения.

Другая программа «Автоматизированный анализ кибербезопасности мобильных приложений» (*Automated Program Analysis for Cybersecurity - APAC)* позволяет осуществлять контроль не декларируемых возможностей (НДВ) в мобильных приложениях в автоматизированном режиме.

Программа (*Integrated Cyber Analysis System - ICAS)* посвящена разработке новых методов автоматического обнаружения и нейтрализации кибератак на основе интеллектуального анализа данных и выявления скрытых закономерностей.

Программа *(Safer Warfighter Computing - SAFER)* предусматривает создание программных средств компьютерной разведки и преодоления средств защиты информации противоборствующей стороны.

В рамках программы Supply Chain Hardware Intercepts for Electronics Defense предполагается разработать миниатюрный (100X100 мкм) и недорогой (меньше одного цента за штуку) чип, который будет подтверждать аутентичность электронных компонентов. Чип будет находиться внутри корпуса микросхемы, но никак не будет электрически связан с ее функциональной начинкой и не должен требовать существенных изменений процесса производства. Ожидается, что эта разработка будет иметь большой успех и на рынке потребительской электроники, где производители не всегда способны проконтролировать качество всех необходимых компонентов (учитывая огромные темпы развития полулегального китайского фабричного производства).

Программа (*Crowd Sourced Formal Verification - CSFV*) направлена на решение сложных аналитических задач в игровой форме. Сложные математические задачи можно представить в виде интересных и увлекательных онлайн-игр.

Программа-конкурс (*Cyber Grand Challenge - CGC*) направлена на разработку приложений для автоматического исправления уязвимостей так называемого нулевого дня, 0-day, в том числе для тестирова-

ния программного обеспечения, выявления уязвимостей, генерации патчей и установки их в компьютерной сети. Сегодня поиск уязвимостей и так частично автоматизирован. Есть методы статичного и динамичного анализа, которые способны обнаружить характерные уязвимости в коде. Но вот исправлять эти ошибки автоматически компьютеры пока не научились. Задача программы - совместить анализ кода и защиту сетей в единый программно-аппаратный комплекс. Задача чрезвычайно сложная, но и 2 миллиона — достойная награда за победу в названной программе-конкурсе. Заметим, что название Cyber Grand Challenge связано с известным конкурсом Grand Challenge, который трижды проводился для автономных транспортных средств в 2004-2007 гг. Десятки автомобилей-роботов пытались проехать по незнакомой пересеченной местности, прокладывая маршрут и объезжая препятствия, включая канавы, камни и узкие тоннели. Маршрут объявляли за два часа до начала конкурса. Если в первый год проведения DARPA Grand Challenge ни одна машина не доехала до финиша (собственно, только 8 из 15 машин смогли уйти со старта, а лучшая команда преодолела 11,8 км из 230), то потом они уже начали соревноваться на скорость. Прогресс был очевиден. По аналогии и с Cyber Grand Challenge — поначалу задача кажется неразрешимой, но в дальнейшем ожидается получить первые самообучающиеся прототипы, которые действительно смогут автоматически генерировать патчи и закрывать уязвимости нулевого дня, 0-day, в течение нескольких секунд после обнаружения. Конкурс Cyber Grand Challenge пройдет в несколько этапов и будет продолжаться несколько лет. В ближайшее время опубликуют информацию о грантах на разработку технологий для проведения этого конкурса, в том числе на создание набора задач. Финал соревнований состоится в первой половине 2016 года.

Программа *Gargoyle* направлена на создание систем киберзащиты, работающих на скоростях более 10 ТБ/сек. Развитие вычислительных возможностей телекоммуникационных средств существенно запаздывает, не справляясь со все возрастающим потоком данных. Результат – пропущенные предупреждения и запоздалая реакция. Например, совокупный мировой поток данных через оптоволоконные кабели в настоящее время составляет более 100 ТБ/сек и, как ожидают, превысит 1 ПБ/sec к 2020 году. В рамках программы Gargoyle в DARPA разрабатывают фотонные корреляторы для обработки критически значимой информации, которые обеспечат почти нулевое время ожидания данных. Для этого создадут широкополосную модуляцию с прямым расширением спектра полосы пропускания более 10 ГГц.

Программа *“Разработки нового беспроводного стандарта связи” (100 Gb/s RF Backbone).* Сейчас в армии США для различного рода коммуникаций применяется безопасный беспроводной протокол Common Data Link (CDL). Он обеспечивает максимальную скорость передачи данных на уровне 250 Мб/с. Однако этой скорости уже недостаточно для полноценного управления беспилотными летательными аппаратами, отправления и получения разведывательных данных. Цель программы 100 Gb/s RF Backbone – создание нового беспроводного стандарта связи, который способен обеспечить скорость передачи данных 100 Гб/с при радиусе покрытия 200 км. При этом требования к массе конечного оборудования и уровню его энергопотребления предъявляются такие же, как и к оборудованию CDL.

В рамках программы *“Разработки новой навигационной системы на смену GPS” (Adaptable Navigation Systems)* разрабатывается система, которая позволит военнослужащему ориентироваться в любых условиях, в том числе тогда, когда сигнал GPS недоступен (например, в результате радиоэлектронного противодействия, особенностей ландшафта и природных явлений). Предполагается, что, во-первых, будет разработан инерциальный измерительный блок нового типа, которому требуется меньше фиксаций координат от системы GPS, например, за счет использования сверхкомпактных атомных часов, работающих с холодными атомами. Во-вторых, будет создан метод использования эфирных сигналов (SoOp) от различных источников – наземного, воздушного и космического базирования. В-третьих, будет улучшена навигация по геофизическим полям. В результате должна появиться новая многоцелевая навигационная система, которая сможет изменять конфигурацию в полевых условиях для работы произвольного оборудования в различных условиях эксплуатации.

Программа *XDATA* направлена на создание систем киберзащиты на основе больших данных, Big Data. В рамках упомянутой программы в открытом каталоге *DARPA* представлены проекты с исходными кодами на *GitHub* под свободными лицензиями: *ALv2, BSD, GPL, GPLv3, LGPL, MIT и др. В* том числе, библиотека на языке Python компании *Continuum Analytics* для интерактивной визуализации «больших данных»; *Bokeh д*ля «тонких» клиентов; библиотека для построения масштабируемых байесовых сетей *SMILE-WIDE* для Boeing с соответствующим API-интерфейсом (представляет собой аналог известного API SMILE, который дополнительно способен исполнять векторные операции за счет распределенной реализации на Hadoop), оптимизирующий компилятор Numba для Python, разработанный Continuum Analytics, Inc. под лицензией BSD и пр.

3. Модель киберугроз. Анализ результатов поисковых исследований агенства DARPA в области кибербезопасности, а также ряда смежных НИОКР свидетельствует о появлении и необходимости нейтрализации новых классов угроз кибербезопасности [4–10]. В том числе, следующих киберугроз цифровой обработки данных:

Угроза подключения к каналам связи.

Цифровая обработка сигналов дает возможность копирования («ответвления») голосового трафика в пределах коммутационной матрицы без каких бы то ни было демаскирующих признаков. Факт копирования невозможно отследить, он не вызывает ни изменений в амплитуде передаваемого сигнала, ни искажений, связанных с задержкой передачи. Это является качественным отличием современных вычислительных систем и сетей.

В частности, практически все крупные разработчики оборудования реализовали в программном обеспечении те или иные возможности копирования речевого трафика при наличии у прослушивающей стороны соответствующих полномочий, определенных администратором. В некоторых случаях это полноценная трехсторонняя конференцсвязь с отключенным входящим голосовым каналом от прослушивающей стороны, в других – ответвление потока по специальной схеме при наборе определенного номера. Некоторые исследователи в области информационной безопасности отдельно выделяют так называемый «полицейский режим» – возможность выполнения тех же операций извне, при наборе из городской телефонной сети определенного номера, принадлежащего номерному полю УАТС, и кода допуска

Цифровые учрежденческие АТС модели AVAYA Definity реализуют возможность скрытого копирования речевой информации в рамках возможности «Service Observing» (Контроль вызова), позиционируемой как средство для контроля со стороны менеджеров за ходом работы телефонных операторов, в первую очередь в Центрах обработки вызовов. Активация функции возможна как в варианте с подачей в речевой канал предупредительного сигнала каждые 12 секунд о факте прослушивания третьей стороной, так и без него. Настройка полномочий на прослушивание выполняется с консоли администратора по групповому принципу: каждой абонентской линии соотносится класс приоритетов «COR», а в матричной форме для каждой пары классов определяется разрешение или запрет прослушивания. Активация прослушивания выполняется набором кода доступа к сервису, а затем номера абонента, и может быть назначена на одну из функциональных клавиш прослушивающего аппарата. Кроме того, при определенной

настройке возможен доступ к функции с внешних линий, например, с городской телефонной сети.

Сервер IP-телефонии CallManager от компании Cisco Systems Inc. также предоставляет возможность включения в разговор третьего абонента, обладающего достаточными полномочиями (как с предупредительным сигналом, так и без него). Функция именуется «Barge In» и имеет две различные схемы технической реализации.

1) Схема на основе программно-аппаратных средств, штатно встроенных во все IP-аппараты компании с 2-мя линиями. Прослушиваемый IP-аппарат при поступлении запроса на конференц-связь (в т.ч. одностороннюю – прослушивание) самостоятельно выполняет ответвление и микширование двух голосовых потоков (первичного – в направлении абонента и вторичного – в направлении прослушивающего устройства) аппаратными средствами второй линии. При этом при соответствующей настройке предупредительных сигналов в первичный голосовой поток не добавляется, более того, на дисплее прослушиваемого IP-аппарата не появляется никаких информационных признаков о факте подключения. Данная схема ограничена только одним подключением прослушивания и только широкополосным (64 кбит/с) кодеком G.711, однако, не вносит никаких демаскирующих искажений в голосовой поток.

2) Схема на основе выделенных программно-аппаратных средств конференц-связи сервера IP-телефонии. При поступлении запроса сервер IP-телефонии замыкает голосовой трафик в обоих направлениях (проходивший до этого момента напрямую между IP-устройствами) на устройство конференц-связи и с его помощью выполняет микширование и ответвление данных (в этом случае уже на неограниченное количество прослушивающих устройств и вне зависимости от используемого абонентами кодека). Недостатком схемы по сравнению с первым вариантом является слышимое искажение («провал голоса») в момент переключения потоков.

Настройка привилегий на прослушивание выполняется отдельно для каждой прослушиваемой линии (непосредственно указывается набор линий, имеющих право на подключение, в т.ч. незаметное, к разговору).

Таким образом, получение противником тем или иным образом привилегий администратора предоставляет ему практически неограниченные возможности по незаметному прослушиванию.

Угроза прослушивания разговоров в помещении с помощью автоответа.

Цифровые и IP- аппараты, как сложные компьютерные устройства, привнесли еще один класс угроз утечки речевой информации, связанный с возможностью удаленного (в т.ч. при некоторых условиях – несанкционированного) включения микрофона и передачи разговоров, ведущихся в помещении по цифровому каналу. В качестве первого, рассмотрим вариант, не связанный с недокументированными возможностями самих аппаратов – широко распространенную опцию «Автоответ». При ее активации вызываемый аппарат при поступлении вызова подает один (часто – укороченный) сигнал вызова, а затем автоматически включает микрофон и громкоговоритель с тем, чтобы абоненты имели возможность общаться между собой по громкой связи либо с использованием гарнитуры.

При возможности настройки опции «Автоответ» в зависимости от вызывающей линии (интерком) она начинает представлять реальную угрозу прослушивания разговоров, ведущихся в помещении. Злоумышленник, получивший привилегии администрирования УАТС, может создать интерком-группу, включив в нее атакуемую линию и свой номер, изменить сигнал вызова со своей линии на запись тишины и получить тем самым возможность прослушивать разговоры в помещении, сделав вызов на данную линию. Схема обладает некоторыми незначительными демаскирующими признаками: 1) в зависимости от модели аппарата факт включения микрофона может отражаться индикаторами, 2) линия в момент прослушивания будет занята при попытке вызова извне, и 3) существует риск поднятия прослушиваемым абонентом трубки для выполнения вызова. Однако, это не исключает возможность выполнения успешного и скрытного прослушивания, особенно в ситуациях, когда в помещении идет активное обсуждение того или иного вопроса, а телефонный аппарат установлен так, что его индикаторы не видны присутствующим.

Угроза наличия недокументированных возможностей.

Недокументированные возможности самих аппаратов (в особенности IP-) являются еще одной угрозой для конфиденциальности речевой информации в защищаемых помещениях. Программное обеспечение IP-телефонов представляет собой сложный программный комплекс, в т.ч. реализующий стек протоколов TCP/IP, и может содержать:

- недокументированные возможности, внесенные разработчиками в целях тестирования или на определенных этапах разработки новых функциональных возможностей аппаратов;

- ошибки в реализации, например, приводящие к уязвимостям класса «переполнение буфера», и позволяющие получить полный контроль над программным обеспечением аппарата до его перезагрузки.

Примером угрозы первой группы является имевшаяся в одной из версии ПО возможность отправки на IP-телефоны наиболее популярных моделей 7940 и 7960 компании Cisco Systems Inc. управляющего XML-сообщения CiscoIPPhoneExecute, которое среди прочих возможностей (набор номера, эмуляция нажатия клавиш и т.п.) могло включать микрофон аппарата и передавать весь голосовой трафик на указанный в XML-сообщении IP-адрес.

Угроза прослушивания IP-трафика в момент передачи по сети.

Различные варианты реализаций угроз прослушивания трафика традиционны для компьютерных сетей, использующих в своей структуре широковещательные сегменты (Ethernet, в т.ч. коммутируемый, радио-Ethernet и т.п.), и создают еще один уровень возможных атак на системы IP-телефонии. При отсутствии шифрования трафика на сетевом или более высоких уровнях модели OSI существует несколько вариантов нарушения конфиденциальности передаваемых сообщений.

В условиях отсутствия у злоумышленника административных прав на активное сетевое оборудование наиболее эффективной в коммутируемых Ethernet-сетях является атака «ARP spoofing», выполняющая изменение таблицы маршрутизации на канальном (MAC) уровне с помощью специально сформированных ARP-пакетов. Также к раскрытию определенной части передаваемой информации может привести перевод коммутатора в режим концентратора с помощью большого количества фальшивых пакетов (MAC storm), хотя этот способ и обладает значительными демаскирующими признаками, выражающимися в резком снижении качества работы сети.

При получении злоумышленником административных прав на коммутирующем или маршрутизирующем оборудовании (например, в результате атаки на компьютер администратора или при перехвате его пароля, передавшегося в открытом виде) у него появляются гораздо более мощные средства перехвата IP-трафика. Они включают:

- возможность активации на коммутаторах зеркальных (SPAN) портов, получающих точную копию передаваемого по определенным портам трафика;
- использование иных технологий «ответвления» трафика от производителей сетевого оборудования, например:
  - протокола ERSPAN (Encapsulated Remote SPAN), инкапсулирующего каждый перехватываемый пакет в пакет протокола

GRE, что позволяет передавать его по IP-сетям без каких-либо ограничений дальности;

– опции IP Traffic Export, реализующей "ответвление" трафика при его маршрутизации на 3-ем уровне модели OSI;

Оба протокола поддерживают возможность тонкой настройки фильтрации перехватываемых пакетов, что позволяет копировать трафик только от определенных групп IP-устройств.

Беспроводные сети при отсутствии стойких алгоритмов шифрования также являются потенциальным источником раскрытия передаваемого по ним голосового трафика.

Угроза подмены сообщений в управляющем канале.

Методика централизованного управления IP-телефонными вызовами (реализуемая в УАТС) содержит еще один возможный путь прозрачного для абонентов перехвата их разговоров. В момент установления IP-соединения первоначальный обмен информацией, содержащей номера абонентов, их имена, технические возможности аппаратов и т.п., в т.ч. IP-адреса оконечных устройств, идет между серверами IP-телефонии. На этом этапе возможна подмена (средствами атак сетевого уровня) информации об одном или обоих IP-адресах с целью внедрения противника в цепочку передачи голосового трафика по принципу прозрачного прокси-сервера.

Подобный класс атак остается совершенно незаметным на прикладном уровне, т.к. пользователю обычно не видны сетевые координаты удаленного абонента, а стек протоколов не способен обнаружить факт подмены, и может быть выявлен только с помощью специализированного мониторинга сетевого трафика.

В целом, предпосылкой для появления возможности подобных атак является то, что в современных протоколах IP-телефонии (H.323, SCCP и др.) оконечное оборудование при приеме и передаче голосового потока является ведомым относительно сервера УАТС и полностью полагается на информацию, сообщенную ему в управляющем канале (в т.ч., например, не проверяет соответствие IP-адресов отправителя и получателя голосового потока в рамках одного и того же разговора). Проблема обеспечения защиты от внедрения в голосовой поток прокси-сервера поднимает вопрос об обеспечении целостности передаваемых в управляющем канале данных стойкими криптографическими методами.

4. Заключение. Анализ результатов поисковых исследований DAPRA [3] свидетельствует о появлении и необходимости нейтрализации новых угроз кибербезопасности. Известные публикации газеты «The Washington Post» на основе заявлений Э.Сноудена подтверждают

реализацию упомянутых киберугроз на практике. Разведслужбы США в течение только 2011 года провели против других стран 231 кибератаку, были потрачены более 652 мл. долларов США. При этом три четверти кибератак были направлены против России, Ирана, Китая и Северной Кореи, в т.ч. ядерные программы этих стран. По мнению Э. Сноудена США провели более 61 тысячи хакерских кибератак по всему миру. Все вместе это подтверждает актуальность постановки и проведения специальных НИОКР по своевременному выявлению и парированию выявленных киберугроз цифровой обработки данных.

Литература

1. *Клабуков И.Д., Алехин М.Д., Нехина А.А.* Исследовательская программа DARPA на 2015 год // М. 2013. 102 с.
2. *Клабуков И.Д., Алехин М.Д., Мусиенко С.В.* Сумма технологий национальной безопасности и развития // М. 2013. 110 с.
3. Официальный сайт агентства по перспективным оборонным научно-исследовательским разработкам Defense Advanced Research Projects Agency, DARPA. URL: www.darpa.mil (дата обращения: 12.01.2015).
4. *Kellerman T.* Cyber-threat proliferation: Today's truly pervasive global epidemic // Security Privacy, IEEE. 2010. vol. 8. no. 3. pp. 70–73.
5. *Wilshusen G.C.* Cyber threats and vulnerabilities place federal systems at risk: Testimony before the subcommittee on government management, organization and procurement // United States Government Accountability Office. Tech. Rep. 2009.
6. *Musliner D.J., Rye J.M., Thomsen D., McDonald D.D., Burstein M.H.* FUZZBUSTER: Towards adaptive immunity from cyber threats // In 1st Awareness Workshop at the Fifth IEEE International Conference on Self-Adaptive and Self-Organizing Systems. 2011. pp. 137–140.
7. *Musliner D.J., Rye J.M., Marble T.* Using concolic testing to refine vulnerability profiles in FUZZBUSTER // In SASO-12: Adaptive Host and Network Security Workshop at the Sixth IEEE International Conference on Self-Adaptive and Self-Organizing Systems. 2012. pp. 9–14.
8. *Musliner D.J., Friedman S.E., Rye J.M., Marble T.* Meta-control for adaptive cybersecurity in FUZZBUSTER // Proc. of 7th IEEE Int. Conf. on Self-Adaptive and Self-Organizing Systems. 2013. pp. 219–226.
9. *Burnim J., Sen K.* Heuristics for scalable dynamic test generation // Proceedings of the 23rd IEEE/ACM International Conference on Automated Software Engineering, ser. ASE '08. 2008, pp. 443–446. URL: http://dx.doi.org/10.1109/ASE.2008.69.
10. *Weimer W., Forrest S., Goues C.Le, Nguyen T.* Automatic program repair with evolutionary computation // Communications of the ACM. 2010. vol. 53. no. 5. pp. 109–116.

References

1. Klabukov I.D., Alehin M.D., Nehina A.A. *Issledovatel'skaja programma DARPA na 2015 god* [DARPA research program for 2015]. M. 2013. 102 p. (In Russ.).
2. Klabukov I.D., Alehin M.D., Nehina A.A. *Summa tehnologij nacional'noj bezopasnosti i razvitija* [Sum of technologies to national security and development]. M. 2013. 110 p. (In Russ.).

3. Oficial'nyj sajt agentstva po perspektivnym oboronnym nauchno-issledovatel'skim razrabotkam [Official web site of Defense Advanced Research Projects Agency, DARPA]. Available at: www.darpa.mil. (accessed 12.01.2015).
4. Kellerman T. Cyber-threat proliferation: Today's truly pervasive global epidemic. *Security Privacy, IEEE*. 2010. vol. 8. no. 3. pp. 70–73.
5. Wilshusen G.C. Cyber threats and vulnerabilities place federal systems at risk: Testimony before the subcommittee on government management, organization and procurement. United States Government Accountability Office. Tech. Rep. 2009.
6. Musliner D.J., Rye J.M., Thomsen D., McDonald D.D., Burstein M.H. FUZZBUSTER: Towards adaptive immunity from cyber threats. In 1st Awareness Workshop at the Fifth IEEE International Conference on Self-Adaptive and Self-Organizing Systems. 2011. pp. 137–140.
7. Musliner D.J., Rye J.M., Marble T. Using concolic testing to refine vulnerability profiles in FUZZBUSTER. In SASO-12: Adaptive Host and Network Security Workshop at the Sixth IEEE International Conference on Self-Adaptive and Self-Organizing Systems. 2012. pp. 9–14.
8. Musliner D.J., Friedman S.E., Rye J.M., Marble T. Meta-control for adaptive cybersecurity in FUZZBUSTER. Proc. of 7th IEEE Int. Conf. on Self-Adaptive and Self-Organizing Systems. 2013. pp. 219–226.
9. Burnim J., Sen K. Heuristics for scalable dynamic test generation. Proceedings of the 23rd IEEE/ACM International Conference on Automated Software Engineering, ser. ASE '08. 2008, pp. 443–446. URL: http://dx.doi.org/10.1109/ASE.2008.69.
10. Weimer W., Forrest S., Goues C.Le, Nguyen T. Automatic program repair with evolutionary computation. Communications of the ACM. 2010. vol. 53. no. 5. pp. 109–116.

Петренко Сергей Анатольевич — д-р техн. наук, директор, Центр систем кибербезопасности АФК "Система", научный эксперт Совета Безопасности РФ. Область научных интересов: компьютерные науки, кибербезопасность, программирование. Число научных публикаций — 80. s.petrenko@rambler.ru, http://www.htsts.ru; Б. Ґрузинская, д. 12, стр. 2, Москва, 123242; р.т.: +79037428543, Факс: +79037428543.

Petrenko Sergei Anatol'evich — Ph.D., Dr. Sci., director, Center for Systems Cybersecurity AFK "Sistema", scientific expert of the Security Council of the Russian Federation. Research interests: computation with memory, cybersecurity, theoretical programming. The number of publications — 80. s.petrenko@rambler.ru, http://www.htsts.ru; 12, Bolshaya Gruzinskaja, Building 2, Moscow, 123242; office phone: +79037428543, Fax: +79037428543.

РЕФЕРАТ

Петренко С.А. **Модель киберугроз по аналитике инноваций DARPA.**

В работе проведен анализ современных поисковых исследований агенства DARPA в области кибербезопасности. Определена новая модель киберугроз цифровой обработки данных. Здесь под киберугрозами понимаются действия или события, которые могут привести к нарушению способности выполнения основных целевых задач объекта информатизации. По характеру проявления киберугрозы разделены на преднамеренные (умышленные) и непреднамеренные (случайные). Косвенные киберугрозы связаны с попытками несанкционированного доступа к информации без физического доступа к элементам системы, а прямые – с попытками несанкционированного доступа к информации с физическим доступом к элементам системы. В зависимости от используемых методов реализации киберугрозы разделены на пассивные и активные. Пассивные киберугрозы имеют целью несанкционированный доступ к информации без изменения состояния информационной системы. Примером пассивной угрозы может служить перехват передаваемой по каналам и линиям связи информации для последующего ее анализа. Активные киберугрозы имеют целью намеренное несанкционированное изменение состояния, например, хищение или модификации конфиденциальной информации. В качестве источника активных угроз могут также выступать специальные средства, называемые закладками (программными, аппаратными или программно-аппаратными), компьютерные вирусы, вредоносное программное обеспечение, которые могут быть встроены в программно-аппаратные средства информационных систем на этапе их изготовления, или внедрены в них в процессе эксплуатации.

Подробно рассмотрены следующие киберугрозы цифровой обработки данных: подключение к каналам связи, прослушивание разговоров в помещении с помощью автоответа, наличие недокументированных возможностей, прослушивание IP-трафика в момент передачи по сети, подмена сообщений в управляющем канале. Определена соответствующая модель киберугроз цифровой обработки данных и предложены рекомендации по своевременной нейтрализации упомянутых киберугроз.

SUMMARY

Petrenko S.A. **Model Cyber Threats by Analysis of DARPA Innovations.**

The analysis of modern exploratory research of agency DARPA in cybersecurity was carried out. A new model of cyber digital data processing is defined. Here the activities under cyber threats or events that may disrupt the ability to perform the basic targets of the object information are analyzed. Cyberthreats by the nature of manifestations are divided into intentional (deliberate) and unintentional (accidental). Indirect cyber threats are associated with unauthorized access to information without having physical access to the elements of the system, and direct - to attempt to gain unauthorized access to information with physical access to the system elements. Depending on the used methods of implementing cyberthreats divided into passive and active. Passive cyberthreats aim to unauthorized access to the information without changing the state of the information system. An example of a passive threat can serve the intercept of transmitted channels and lines of communication of information for its subsequent analysis. Active cyberthreats aim intentional unauthorized changes to the state, such as theft or modification of sensitive information. As a source of active threats may be also special tools called tabs (software, hardware, or software and hardware), computer viruses, malicious software, which can be embedded in software and hardware information systems at the stage of manufacture, or embedded in them during operation.

Discussed in detail the following cyberthreats for digital data processing: connection to communication channels, listening to conversations in the room with the help of auto-answer, the presence of undocumented features, listening to IP-traffic at the time of transmission over the network, the substitution of messages in the control channel. The appropriate model cyberthreats of digital data and recommendations on timely neutralizing mentioned cyberthreats are determined.

УДК 004.77; 004.94; 510.635.3

В.А. Овчаров

МОДЕЛИРОВАНИЕ СУБЪЕКТНО-ОБЪЕКТНОГО ВЗАИМОДЕЙСТВИЯ В СЕТЕВЫХ ИНФРАСТРУКТУРАХ

Овчаров В.А. **Моделирование субъектно-объектного взаимодействия в сетевых инфраструктурах.**

Аннотация. В работе рассматривается задача идентификации различных аспектов функционирования взаимодействующих объектов информационно-телекоммуникационных сетей (ИТКС) по результатам мониторинга сетевого трафика. В качестве решения данной задачи в части идентификации типов сетевых объектов и операций взаимодействия предлагается графовая модель поведения, в части деанонимизации отношений взаимодействующих объектов предложены предикатные модели состояний объектов информационно-телекоммуникационной сети (ИТКС) на основе отношений между экземплярами.

Ключевые слова: средства активной идентификации, средства пассивной идентификации, мониторинг сетевого трафика, сетевой процесс, контроль поведения, сетевой объект, информационно-телекоммуникационная сеть.

Ovcharov V.A. **Simulation of Subject-Object Interaction in Network Infrastructures.**

Abstract. The problem of identification of the different aspects of the interacting objects information and telecommunication networks (ITN) on the results of network traffic monitoring is analyzed. As a solution to this problem in terms of identifying the types of network facilities and operations interaction graph model of behavior is proposed, in terms of relations disclosure of anonymity interacting objects predicate state model objects ITN based on the relationship between instances is offered.

Keywords: active means of identification means of passive identification, network traffic monitoring, network process, behavior control, network object information and telecommunications network.

1. Введение. Современные ИТКС представляют собой сложные организационно-технические системы, состоящие из большого числа компонентов различной степени автономности, в разноплановой программно-аппаратной конфигурации, связанных между собой различными по используемым технологиям и скорости передачи данных каналами связи, обменивающиеся данными различных типов. Обеспечение сетевой безопасности и эффективного расследования компьютерных инцидентов, ставших реалиями современных ИТКС различного назначения, немыслимо без использования специалистами по сетевой криминалистике различных систем мониторинга, автоматизирующих процессы первичной обработки сетевого трафика и последующего анализа разнородной технической информации из различных источников.

Одной из наиболее актуальных задач перспективных систем мониторинга и обеспечения комплексной защиты информации в ИТКС является контроль действий приложений на сетевых узлах, сопостав-

ление идентифицированных процессов их взаимодействия с реальными пользователями и построение причинно-следственных связей для прогнозирования дальнейших действий пользователей.

В то же время разрабатываемые системы должны быть универсальными, обеспечивая работу как с различными источниками данных, так и в представленных на рисунке 1 сценариях (1 – сценарий сбора трафика на зеркалируемом интерфейсе, 2 – сценарий мониторинга количества и программно-аппаратных характеристик устройств, находящихся за NAT-/PAT-устройствами и межсетевыми экранами, 3 – сценарий мониторинга интерфейсов беспроводных сетей передачи данных и локальных беспроводных сегментов, 4 – сценарий мониторинга в качестве клиента проводного сегмента сети) использования в ИТКС различного типа. Данные сценарии обуславливают необходимость применения разрабатываемых моделей в ИТКС различного назначения и топологии, учета факторов большого количества узлов, программно-аппаратной неоднородности, динамики, существенной территориальной распределенности.

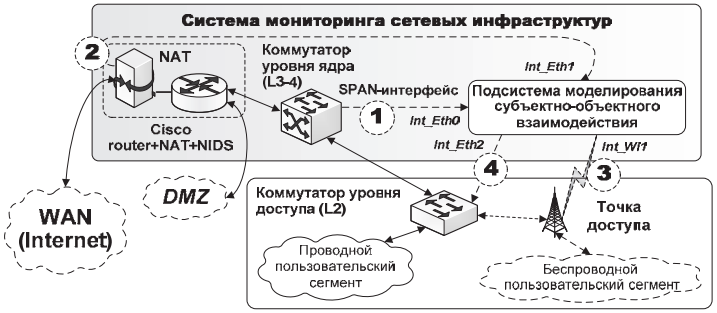


Рис. 1. Сценарии использования перспективных многофункциональных систем мониторинга ИТКС

Таким образом, при разработке соответствующих моделей (многомодельных комплексов) целесообразно употреблять термин «сетевая инфраструктура», подразумевая под ним возможность использования предлагаемых моделей в сетях различной архитектуры (одноранговой и клиент-серверной), типа (проводных IEEE 802.3 и беспроводных IEEE 802.11, 802.16), различных смешанных топологий [19]. В общем случае сетевая инфраструктура представляет собой граф:

$$G = (V, E),$$

где V – множество вершин (сетевых объектов), E – множество связей между сетевыми объектами.

Каждый сетевой объект V представляет собой структуру вида $\{x_1(t), x_2(t), \ldots, x_n(t)\}$, где $x_i(t)$ – параметр объекта, определенный в момент времени t и который может быть вычислен на основе обработки результатов мониторинга сетевого трафика. Таким образом, задача идентификации параметров сетевых объектов формулируется следующим образом: для каждого v из V определить вектор параметров v' и определить, насколько v' соответствует типовому профилю сетевого объекта из базы профилей.

Для решения данной задачи предлагается следующая функциональная декомпозиция. Концептуальная модель субъектно-объектного взаимодействия, учитывающая логические связи реальных пользователей ИТКС, ассоциированных носителей действий данных пользователей (набора идентификаторов, связанных с субъектом) и сетевыми объектами, идентифицируемыми средствами пассивного анализа сетевого трафика. Для выделения инициаторов информационного обмена разработана графовая модель поведения. Для формирования выводов по результатам анализа изменения состояний взаимодействующих экземпляров предложены предикатные модели пассивных и активных объектов ИТКС. В качестве подхода к решению проблемы наблюдаемости объектов ИТКС в информационном пространстве разработана модель ассоциированного представления процессов взаимодействия «субъект-различные типы устройств».

Специфика решаемой в работе задачи потребовала на первом этапе уточнения некоторых терминов и определений из работ [15, 16, 21], которые представлены следующим образом.

Мониторинг сетевого трафика – процесс систематического наблюдения за объектами и субъектами, влияющими на безопасность исследуемой ИТКС, сбора информации о параметрах состояния ее программно-аппаратных средств, социально-коммуникационных аспектах взаимодействия субъектов, сетевых и физических объектов, а также анализа и обобщения результатов наблюдений с целью фиксации соответствия (несоответствия) результатов первоначальным предположениям.

Сетевой процесс представляет собой профиль такого поведения ИТКС (динамической системы), которое заключается в исполнении действий по приему или передаче пакетов сетевого трафика или их преобразованию.

Коммуникационный протокол – совокупность правил, регламентирующих формат и процедуры обмена информацией между двумя или несколькими независимыми устройствами, компьютерами, программами или процессами.

Объект ИТКС – одна из сторон взаимодействия в ИТКС, ассоциированная с одним или несколькими субъектами, способная инициировать выполнение операций в соответствии с определенным протоколом.

Субъект ИТКС – ассоциированный с реальным пользователем носитель действий, поведение которого регламентируется политикой безопасности ИТКС или правилами разграничения доступа.

Пассивный мониторинг сетевых инфраструктур – комплекс технических мероприятий по сбору информации для формирования коммуникационных и поведенческих портретов объектов ИТКС на основе сбора, первичного анализа и декодирования сетевого трафика, а также информации уровня приложений на выбранном интерфейсе (группе интерфейсов), подмены и манипуляции данными без использования механизмов установления транспортных соединений.

Активный мониторинг сетевых инфраструктур – комплекс технических мероприятий по сбору информации для формирования коммуникационных и поведенческих портретов объектов ИТКС на основе анализа информации о состоянии коммуникационных портов, запущенных сервисах, службах, уязвимостях системного и прикладного ПО, доступности ресурсов, технологических процессах (циклах) функционирования, связанных с их работой, использующий механизмы установления транспортных соединений.

Коммуникационный портрет сетевых инфраструктур – форма описания и отображения характеристик сетевой инфраструктуры в целом и объекта ИТКС в части детализации правил (коммуникационных протоколов) и процедур обмена данными на соответствующем интервале наблюдения.

Поведенческий портрет сетевых инфраструктур – характерный индивидуальный штамп (параметрическое множество), характеризующий сферу профессиональных и личных интересов, кругозор, опыт, круг общения, образ и ритм жизни пользователей ИТКС.

2. Анализ существующих подходов к разработке моделей представления процессов функционирования сетевых инфраструктур. Вопросы сетевой анонимности в распределенных системах и идентификации злоумышленников на основе анализа трафика Tor-клиентов активно рассматриваются в работе [7]. Недостатком предложенных моделей и метрик является необходимость обеспечения доступа к магистральным каналам провайдеров уровня Tier-1, в то время как корректность работы предложенного симулятора TorPS и достоверность полученных результатов оценить невозможно из-за отсутствия доступа к приложению и его исходному коду.

Подавляющее большинство известных систем активного и пассивного мониторинга используют неформальные модели теории параметрической идентификации для выявления вредоносных http-запросов [13], а также сигнатурные методы [10], для которых на данный момент сложно получить теоретические оценки полноты, показать корректность, завершаемость [18, 4].

Средства пассивной идентификации (p0f, satori, network miner, caploader, Ettercap и др.) операционных систем (ОС) и компонент программного обеспечения (ПО) на основе анализа сетевого трафика и загружаемых дампов в различных форматах [19, 8, 12] используют модель вида

$$OSDef = \langle W, T, D, S, O, Q, OS, Det \rangle,$$

где W – значение поля WS (S_{nn} – кратный значению MSS, T_{nn} – кратный MTU), T – значение поля TTL, D – значение поля фрагментации, S – значение поля SYN, O – значение полей опций (N – NOP, E – EOL, W_{nnn} – значение опции масштабирования окна), M_{nnn} – максимальный размер сегмента, S – selective ACK OK, T – временная метка, T_0 – временная метка с нулевым значением), Q – опции (P – options past EOL, Z – значение поля IP ID, I – указанные параметры IP, U – URG-указатель, X – неиспользуемое ненулевое поле, A – число ACK, T – 2-я метка timestamp, F – различные нестандартные флаги (PUSH, URG и др.), D – данные полезной нагрузки, $!$ – опции (значения) некорректного сегмента, OS – тип ОС, Det – описание версии ОС.

В работе [17] разработаны технологии обнаружения и идентификации вредоносных программ на основе методов интеллектуального анализа данных. Выделены основные группы сущностей, используемых для формирования типовых методик обнаружения вредоносных программ на основе данной группы методов. Приведен обзор существующих методик обнаружения вредоносных программ на основе выделенных групп признаков, наборов данных, методов выделения значимых признаков и обучения, а также программных средств поддержки вычислений. Ограничением данных технологий является необходимость привлечения экспертов для разработки адекватных моделей процессов обучения и функционирования таких систем.

Средства активной идентификации ОС и ПО, сканеры портов и уязвимостей типа nmap, nessus, maxpatrol, используют модель вида [1]

$$OSDef = \langle W, M, T, Ws, S, N, D, T, F_n, L, OS \rangle,$$

где W – значение поля размера окна TCP Window Size, M – значение поля размера сегмента TCP Maximum Segment Size, T – значение поля

времени жизни TTL, *Ws* – значение опции масштабирования Window Scale, *D* – значение поля фрагментации, *S* – индикатор опции TCP SACK, *N* – индикатор опции TCP NOP, *D* – индикатор флага IP Don't Fragment, *T* – индикатор временной метки TCP Timestamp, F_n – индикатор флага пакета ($F_S = SYN$, $F_A = SYN+ACK$), *L* – значение длины пакета, *OS* – тип ОС.

Подход к построению систем анализа защищенности на основе активных методов, предложенный в работе [9] базируется на механизме автоматического генерирования и выполнения распределенных сценариев атак с учетом разнообразия целей и уровня знаний злоумышленника. В основе рассматриваемого подхода – комплексное использование основанных на экспертных знаниях моделей злоумышленника, вероятностных моделей ИТКС, генерации комплекса сценариев атак и оценки уровня защищенности.

В работе [1] был предложен, а в [6] – практически апробирован подход к идентификации компонент ПО на основе динамического сопоставления наблюдаемого поведения ОС или приложения с моделью его нормального поведения в виде альтернирующего автомата. В работах [12, 18] рассматриваются вопросы автоматического поиска уязвимостей и верификации протоколов и приложений с использованием эмпирических критериев допустимого поведения системы. Разработанные модели полны с точки зрения описания возможных отказов системы. В то же время, сложность [12] и узкая специализация сценариев использования предложенных в [11] решений позволяет эффективно использовать их на практике только в задачах аудита безопасности или отложенного расследования инцидентов.

Проведенный анализ [2, 3, 4, 14, 20] показал, что методы теории взаимодействующих последовательных процессов позволяют анализировать с приемлемой сложностью модели с очень большим и даже бесконечным множеством состояний. Применение научно-методического аппарата данной теории к задачам анализа сетевого трафика и расследования компьютерных инцидентов (сетевой криминалистики) возможно, в частности, благодаря разработанной в теории процессов технике символьных преобразований выражений, описывающих процессы.

Разработанные и описанные в данной работе модели проектировались как модели предметной области: сбор информации об объектах мониторинга целевой ИТКС средствами пассивной идентификации, для последующего обоснования корректности и оценки вычислительной сложности методов анализа разнородной технической информации из различных источников.

3. Концептуальная модель субъектно-объектного взаимодействия в ИТКС. Представим исследуемую ИТКС как множество взаимодействующих экземпляров типов объектов. Будем выделять 2 типа объектов ИТКС: программные и аппаратные.

Программные объекты ИТКС разделим на *одноузловые* (экземпляры ОС, ПО на узлах ИТКС, взаимодействующие в рамках одного узла (локально)) и *многоузловые* (экземпляры ПО на узлах ИТКС и внешних ресурсах, взаимодействующие с различными узлами ИТКС).

Аппаратные объекты ИТКС включают в себя активное сетевое оборудование (коммутаторы, маршрутизаторы, точки доступа, шлюзы, межсетевые экраны (МСЭ)), серверное оборудование, различные датчики и сенсоры систем обнаружения атак и мониторинга (DPI), пассивное сетевое оборудование (каналообразующая аппаратура) и радиомодемы, локальные ПЭВМ, физические каналы связи между узлами ИТКС (ПЭВМ), мобильные персональные устройства (МПУ).

Множество программных и аппаратных объектов некоторой ИТКС могут взаимодействовать друг с другом (внутри ИТКС), с аналогичными объектами других ИТКС (например, используя инфраструктуру VPN), или внешними объектами (например, web- или DNS-серверами глобальных сетей). С каждым объектом ИТКС может быть однозначно ассоциирован некоторый субъект, структура которого будет подробно рассмотрена в п.6. В общем случае субъекты могут быть ассоциированы (быть наблюдаемы) как с объектами целевой ИТКС (объекта мониторинга), так и с объектами (как внутренними, так и внешними) других ИТКС в соответствии с особенностями информационного обмена объектов целевой ИТКС.

Концептуальная модель субъектно-объектного взаимодействия в ИТКС представлена на рисунке 2. В разработанной модели приняты следующие обозначения: $T^{(1)}$ – цель мониторинга (подсеть, отдельный сегмент ИТКС, распределенная ИТКС, web-ресурс и пр.), $\{U_{Ti}\}$ – множество реальных (физических) пользователей, ассоциированных с целью мониторинга, $\{Obj^{(1)}\}$ – множество объектов цели (ИТКС), идентифицированных системой мониторинга, $T^{(2)}$ – сегмент ИТКС аналогичного объекта (внутренняя подсеть или ресурс), $T^{(3)}$ – внешний ресурс в глобальной сети или территориально-распределенная (многосегментная, многофилиальная) ИТКС, $Res_{loc_1}^{(1)}$ – внутренние ресурсы цели мониторинга, Res^{ext} – внешние ресурсы, связанные в объектами ИТКС (цели мониторинга), $Obj^{1)}$ – наблюдаемые объекты ИТКС $T^{(1)}$, взаимодействующие в рамках политики безопасности и коммуникаци-

онных протоколов, $Subj^{(1)}$ – идентифицированные в процессе мониторинга субъекты ИТКС $T^{(1)}$.

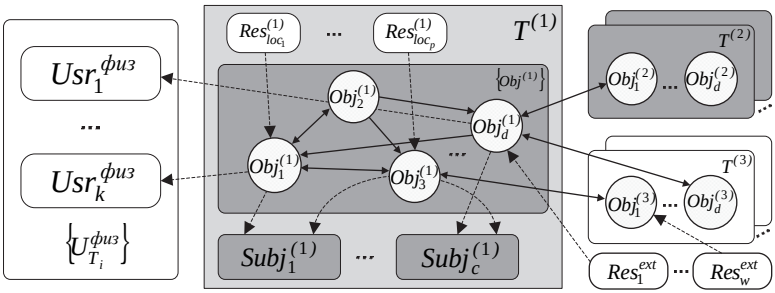


Рис. 2. Концептуальная модель субъектно-объектного взаимодействия в ИТКС

Интерпретация субъекта в соответствии с п.3 связана с тем, что реальный пользователь не имеет доступа к сетевым объектам напрямую, а осуществляет его, используя различные идентификаторы, учетные записи и соответствующие процедуры аутентификации и авторизации. Именно субъекты ИТКС в результате анализа и обобщения результатов мониторинга должны быть корректно сопоставлены с реальными пользователями ИТКС.

4. Графовая модель поведения объектов мониторинга. Разделим множество объектов ИТКС на два подмножества: подмножество активных и подмножество пассивных объектов. Пассивный объект ИТКС не может в данный момент времени осуществлять доступ (инициировать TCP-/UDP-взаимодействие) к другим объектам, в то время как активный – может.

Введем следующие обозначения: A – множество типов активных объектов ИТКС, R_c – множество экземпляров активных объектов ИТКС, P – множество типов пассивных объектов ИТКС, R_p – множество экземпляров пассивных объектов ИТКС, тогда

$$R = R_c \cup R_p .$$

Полагаем $R_c \cap R_p = 0$.

Для каждого типа объекта ИТКС $r_i \in R$ определены соответствующие операции доступа с использованием определенного типа протокола и номера TCP-/UDP-порта (прием/отправка e-mail, обновление ОС и прикладного ПО, чтение/запись на удаленном сервере, использование web-браузера, файловый обмен, запуск на исполнение и т.д.). При этом, доступ к соответствующему объекту ИТКС может быть как

непосредственным, так и косвенным – через другие объекты. Доступ по некоторой операции к объекту ИТКС подразумевает выполнение последовательности элементарных операций в соответствии с логикой работы используемого протокола. Предполагаем, что в ходе реализации доступа объектов в исследуемой ИТКС элементарные операции в рамках коммуникационных протоколов выполняются мгновенно, но, в ряде случаев, могут отстоять друг от друга во времени. Данная ситуация обусловлена временными интервалами мониторинга, в которых возможно несколько элементарных операций, ассоциированных с одним объектом ИТКС (параллельное или последовательное обращение к одному или нескольким объектам).

Пусть r – экземпляр объекта ИТКС типа *type*, представляемый пятеркой вида

$$r = \langle ID, type, \xi, O, T, \rho, \iota \rangle,$$

где *ID* – идентификатор (имя экземпляра) объекта ИТКС, *type* – тип объекта, где $type \in A \cup P$, ξ – значение показателя скомпрометированности экземпляра объекта ИТКС, *O* – множество операций взаимодействия, определенных (идентифицируемых системой мониторинга) для данного типа объектов, *T* – значение порога компрометации объектов ИТКС данного типа, ρ – значение показателя доступности объекта ИТКС, ι – индикатор протокола взаимодействия, причем

$$O = \{o_1, o_2, o_3, o_4, o_5, o_6, o_7, o_8, o_9\},$$

где o_1 – использование служб фоновой передачи (BITS) для обновления, o_2 – аутентификация, o_3 – запрос характеристик ПЭВМ объектом назначения, o_4 – VPN-соединение, o_5 – SSL-соединение, o_6 – VoIP-взаимодействие, o_7 – DNS-взаимодействие с использованием DNSSEC, o_8 – RDP-взаимодействие, o_9 – TOR-взаимодействие.

Представленная пятерка формирует поведенческий портрет каждого экземпляра объекта ИТКС. Графовая модель поведения объектов мониторинга на множестве экземпляров объектов ИТКС в некотором ее состоянии представлена на рисунке 3.

Определим ξ и T как переменные на R со значениями в [0, 1], ι – как переменную со значениями [0, 255], соответствующими коммуникационному протоколу, с использованием которого осуществляется взаимодействие между объектами ИТКС, а ρ – как переменную со значениями [0, 65535], соответствующими количеству портов TCP/UDP, с использованием которых осуществляется взаимодействие

между объектами ИТКС. Будем понимать под значением $\xi(r)$ текущее состояние (не скомпрометирован, скомпрометирован) объекта $r \in R$. Считаем, что если $\xi(r) > T(r)$, то поведение объекта $r \in R$ не определено, а его состояние опасное (в случае мониторинга защищаемой ИТКС), или атакуемое (в случае мониторинга целевой ИТКС).

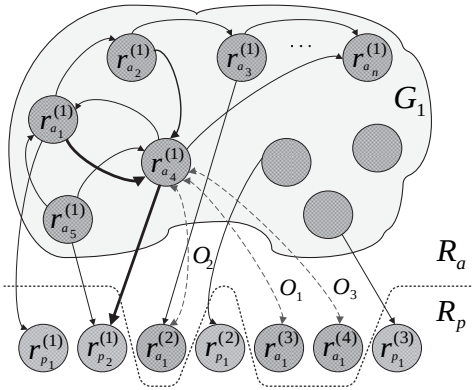


Рис. 3. Графовая модель поведения объектов мониторинга в некотором состоянии ИТКС

Следует отметить, что подсистема визуализации разработанных моделей должна разрешать основное противоречие познания – между *сукцессивностью* (некоторой последовательностью) представления результатов мониторинга и *симультанностью* (одновременностью) их понимания и интерпретации аналитиком. Отсутствие подходов к его разрешению в существующих программно-аппаратных комплексах мониторинга делает их использование неэффективным в повседневной работе аналитика при решении нетривиальных задач и принятии решений. В этой связи необходимо разработать такую подсистему, целевое использование которой позволит оперативно идентифицировать определенные аспекты взаимодействия объектов мониторинга исключительно на основе технологий пассивного анализа сетевого трафика, используя предложенные выше модели. Так, представленная ниже графовая модель (карта ИТКС) позволяет уже на первом этапе разложения графа на плоскости в соответствии с используемым алгоритмом выделить основных инициаторов информационного обмена и их географическое местоположение.

Исходя из приведенных соображений и обозначенных в п.1 сценариев применения систем мониторинга при разработке подсистемы визуализации результатов анализа взаимодействия объектов ИТКС

использовались визуальные средства на основе следующих критериев: разработаны на языке C++, открытый исходный код, мультиплатформенность, лицензия GPLv3. Функциональное представление компонента визуализации (плагина) – кнопка на панели инструментов, при нажатии на которую графовая модель (граф) в соответствующем диалоге изменяется по следующим параметрам: уменьшается число пересечений дуг, среднее расстояние между узлами, дуги могут быть представлены ортогонально, в виде прямых линий. На рисунке 4 приведен пример разложения графа на плоскости (построения карты сети по результатам анализа взаимодействия объектов ИТКС) с использованием разработанного компонента визуализации (библиотеки).

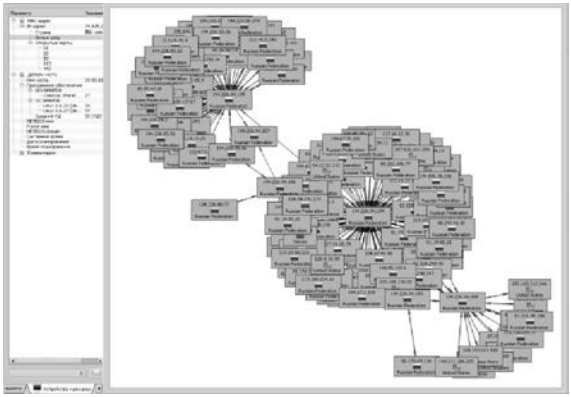


Рис. 4. Пример разложения графа на плоскости (построение карты сети по результатам анализа взаимодействия объектов ИТКС) с использованием разработанного компонента визуализации

Библиотека включает базовые структуры данных, различные классы графов и алгоритмов разложения, которые можно настраивать и изменять с использованием встроенных модулей. В базовые структуры данных, кроме массивов, строк, очередей входят элементы для параллельного программирования (мьютекс, барьер, критическая секция, поток).

5. Предикатные модели состояний объектов ИТКС на основе отношений между экземплярами. Пусть для каждой пары экземпляров объектов ИТКС (r_i, r_j), где $r_i \in R_c, r_j \in R$ системой мониторинга идентифицирована операция взаимодействия $a \in A_{r_j}$, связанная с обращением объекта r_i к r_j с использованием соответствующего про-

токола и порта, тогда предикатная модель (предикат), определяющая отношение взаимодействующих объектов будет иметь вид:

$$a(r_i,r_j) \equiv 1 \Leftrightarrow a \in A_{r_j}\, u(a,r_i,r_j) \in \psi(r_i,r_j),$$

где $\psi(r_i,r_j) = \left\{(a,r_i,r_j) \mid a \in A_{r_j}, r_i \in R_c, r_j \in R\right\}$ – отношение инициализации сессии при взаимодействии объектов r_i и r_j.

Замечание: отношение взаимодействующих объектов не симметрично по $r \in R$ и не транзитивно.

Обозначим тройку $(a,r_i,r_j) \in \psi(r_i,r_j)$ как $\psi_a(r_i,r_j)$, так как это отношение на $R_c \times R$.

В реальных ИТКС пользователи часто используют средства анонимизации типа TOR для разрыва причинно-следственных связей и затруднения работы систем мониторинга. В таких ситуациях взаимодействие объектов может быть не только непосредственным, но и транзитивным, когда один или несколько объектов используют несколько других объектов через цепочку операций взаимодействия. Другим примером является ситуация, когда удаленный пользователь при обращении к терминальному ssh-серверу запускает на выполнение экземпляр сервера, который, в свою очередь, запускает на выполнение определенную программную оболочку или среду виртуальных машин (ВМ), которая, также по командам от пользователя, инициирует новые процессы. При этом все процессы и экземпляры ВМ будут трактоваться как пассивные объекты ИТКС. Последовательность объектов от клиента до конечного пассивного объекта ИТКС при выполнении соответствующей операции будет транзитивным замыканием операции.

Определим подмножество множества активных объектов ИТКС, для которых отношение инициализации сессии по какой-либо операции транзитивно:

$$R_c^{G_1} = \left\{ \begin{array}{l} r_{a_4}^{(1)} \in R\ \exists a : \forall r_{a_1}^{(1)}, r_{p_2}^{(1)} \in R : r_{a_1}^{(1)} \to r_{a_4}^{(1)}, r_{a_4}^{(1)} \to r_{p_2}^{(1)}, \\ r_{a_4}^{(1)} \in R\ \exists a : \forall r_{a_2}^{(1)}, r_{p_2}^{(1)} \in R : r_{a_2}^{(1)} \to r_{a_4}^{(1)}, r_{a_4}^{(1)} \to r_{p_2}^{(1)}, \end{array} \right\}$$

Представим состояние экземпляра пассивного объекта ИТКС r_p тройкой вида:

$$S_{r_p} = (In(r), I(\xi(r_p)), \rho(r_p), \iota(r_p), \vartheta(r_p), \theta(r_p)),$$

где $In(r) = \left\{r \in R_c \mid \exists a \in A_{r_p} : r \to r_p\right\}$ – множество экземпляров объектов, инициирующих и осуществляющих взаимодействие с r_p, $I(\xi(r_p)) = [0,1]$ –

индикатор скомпрометированности экземпляра пассивного объекта ИТКС, $\rho(r_p) = [0,65535]$ – индикатор TCP/UDP-взаимодействия (порта) экземпляра пассивного объекта ИТКС, $\iota(r_p) = [0,255]$ – индикатор протокола взаимодействия (на основе значений поля IPProto) экземпляра пассивного объекта ИТКС, $\vartheta(r_p) = [0,281\ 474\ 976\ 710\ 655]$ – адресный индикатор, $\theta(r_p)$ – DNS-индикатор (доменный индикатор).

Аналогично представим состояние экземпляра активного объекта ИТКС r_c:

$$S_{r_c} = (In(r_c), From(r_c), I(\xi(r_c)), \rho(r_c), \iota(r_c), \vartheta(r_c), \theta(r_c)),$$

где $In(r_c) = \{r \in R_c \mid \exists a \in A_{r_c} : r \to r_c\}$ – множество экземпляров активных объектов ИТКС, инициирующих и осуществляющих непосредственное взаимодействие с r_c, $From(r_c) = \{r \in R \mid \exists a \in A_{r_c} : r_c \to r\}$ – множество экземпляров объектов ИТКС, к которым r_c осуществляет непосредственный доступ и взаимодействие, $I(\xi(r_c)) = [0,1]$ – индикатор скомпрометированности экземпляра активного объекта ИТКС, $\rho(r_c) = [0,65535]$ – индикатор TCP/UDP-взаимодействия (порта) экземпляра активного объекта ИТКС, $\iota(r_c) = [0,255]$ – индикатор протокола взаимодействия (на основе значений поля IPProto) экземпляра активного объекта ИТКС, $\vartheta(r_c) = [0,281\ 474\ 976\ 710\ 655]$ – адресный индикатор, $\theta(r_c)$ – DNS-индикатор (доменный индикатор).

Представленные выражения для состояний экземпляров активных и пассивных объектов ИТКС позволяют сделать вывод о том, что при инициализации соединения и осуществлении дальнейшего взаимодействия некоторого экземпляра объекта ИТКС с соответствующим экземпляром другого объекта изменяются состояния обоих взаимодействующих экземпляров. Таким образом, основной функцией подсистемы анализа аспектов субъектно-объектного взаимодействия является построение дискретно-временной шкалы взаимодействия экземпляров объектов и динамически обновляемой объектной модели данных, отражающей специфику их связей в виде множества узлов и ребер. Узлы и ребра имеют конечный набор атрибутов-идентификаторов: индикатор порта p, протокола l, DNS-имени θ, IP-адреса ϑ. Вся последующая этапу первичной обработки работа системы мониторинга осуществляется с синтезированными или динамически обновленными по результатам первичной обработки предикатными моделями.

6. Модели ассоциированного представления субъектно-объектного взаимодействия. Проблема наблюдаемости объектов ИТКС в информационном пространстве имеет двоякий характер. С одной стороны, она обусловлена тем, что в настоящее время практически каждый реальный пользователь ИТКС использует в различных сетях различные типы устройств: в локальном сегменте ИТКС организации – инфраструктуру терминальных серверов и «тонких клиентов», стационарные ПЭВМ, серверы (рабочие станции) или мобильные рабочие места (МРМ) типа ноутбук, дома, как правило, МРМ, ПЭВМ или мобильной персональное устройство (МПУ), в общественных местах и при перемещениях – МПУ или МРМ. Таким образом, с реальным пользователем (человеком) должны быть корректно ассоциированы все используемые им устройства (типы устройств). Модель ассоциированного представления взаимодействия «субъект-типы устройств» (рисунок 5), призвана решить данный аспект проблемы наблюдаемости объектов ИТКС.

В приведенной модели приняты следующие обозначения: ID^A – персональный идентификатор, ID^B – идентификатор должности в организации (должностной идентификатор), ID^C – идентификатор номера рабочего телефона/факса, ID^D – аппаратный идентификатор (MAC-адрес, IMEI, IMSI, IMN, FHIN, RHIN), ID^F – логический идентификатор (ID VK, e-mail, номер телефона, login), ID^G – идентификатор социальной (тематической) группы, ID^H – AD-идентификатор, NIC_1,…, NIC_J – сетевое имя-псевдоним персоны.

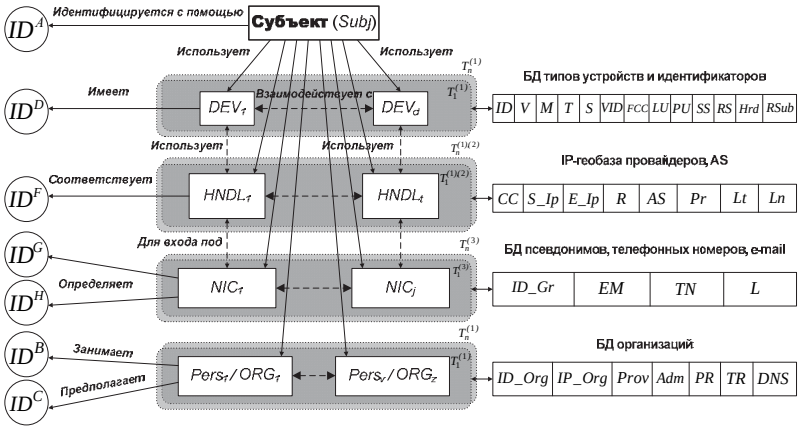


Рис. 5. Схема модели ассоциированного представления процессов взаимодействия «субъект-различные типы устройств»

Столбцы таблиц базы данных (БД) имеют следующие обозначения: *V* – производитель, *M* – модель, *T* – тип, *S* – стандарт, *VID* – идентификатор производителя, *FCC* – FCC-идентификатор, *LU* – login пользователя, *PU* – пароль пользователя, *SS* – тип и версия ПО, *RS* – используемые технологии шифрования, *Hrd* – другие особенности аппаратной части, *Rsub* – особенности радиоподсистемы, *CC* – код страны, *S_Ip* – начальный IP-адрес диапазона, *E_Ip* – конечный IP-адрес диапазона, *R* – идентификатор региона, *AS* – номер автономной системы, *Pr* – наименование провайдера, *Lt* – широта, *Ln* – долгота, *ID_Gr* – идентификатор группы, *EM* – адрес электронной почты, *TN* – номер телефона, *L* – login, *ID_Org* – идентификатор организации, *IP_Org* – IP-адреса организации, *Prov* – реквизиты провайдера, *Adm* – реквизиты администраторов (группы технической поддержки), *PR* – значение page rank, *TR* – значение traffic rank, *DNS* – DNS-имя.

С другой стороны, в реальных ситуациях взаимодействующие объекты мониторинга могут находиться как одном или в различных сегментах целевой ИТКС, так и в различных, территориально распределенных ИТКС. Для решения данного аспекта проблемы наблюдаемости объектов ИТКС разработана модель представления процессов мониторинга, обработки и хранения результатов обработки, формализуемая четверкой вида:

$$M_i: \left\langle S_k, D_l^{dest}, C_m^{dest}, I_p \right\rangle,$$

где S_k – источник информации (объект мониторинга), $k \in N_k$; D_l^{dest} – получатели информации (средства мониторинга (наблюдения)), $d \in N_d$; C_m^{dest} – средства хранения информации, $c \in N_c$; I_p – информационные каналы (каналообразующие средства).

На основе анализа решаемых задач, архитектуры, особенностей программной реализации и специфики взаимодействия функциональных подсистем современных систем мониторинга ИТКС предложена таксономическая схема процессов сбора, обработки и хранения данных мониторинга (рисунок 6), позволяющая синтезировать структурные логико-графические модели систем мониторинга ИТКС различной архитектуры и назначения.

Автоматизация построения подобных моделей представляет собой трудно формализуемую задачу, решаемую, как правило, исследователем эвристически на основе последовательного анализа структуры системы «сверху–вниз». Построение адекватных реальным системам моделей возможно только на основе глубокого знания структуры системы и отдельных подсистем, особенностей функционирования, спе-

цифики информационных потоков и условий эксплуатации. Кроме того, необходимо совершенно чётко представлять возможности обеспечивающих подсистем (резервирования, восстановления и др.) применительно к каждому элементу системы.

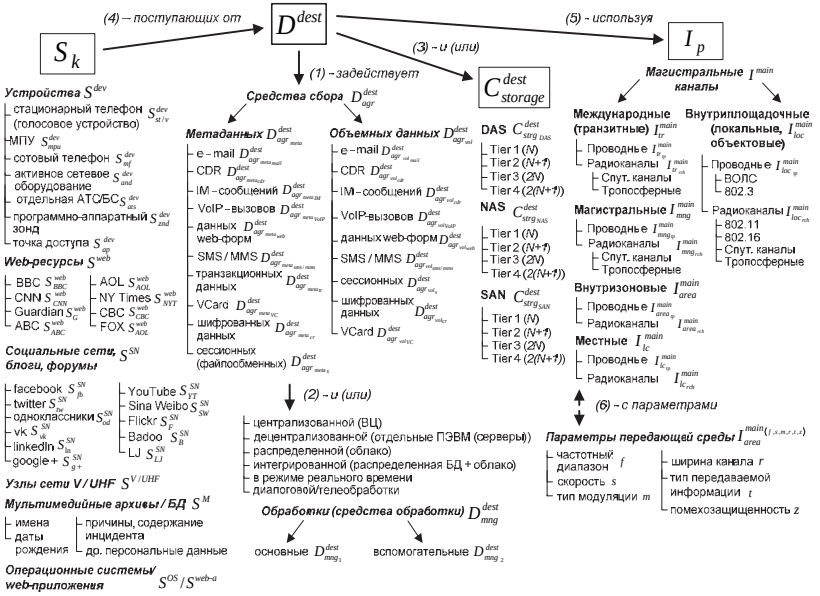


Рис. 6. Таксономическая схема процессов сбора, обработки и хранения данных мониторинга

В соответствии с введенной таксономической схемой типовая модель процессов взаимодействия компонент распределенной системы мониторинга ИТКС (коммуникационный портрет) имеет вид:

$$M : \left\langle \begin{Bmatrix} s^{dev}_{st/v} \cup s^{dev}_{mpu} \cup s^{dev}_{ats} \cup s^{dev}_{mf} \cup s^{dev}_{znd} \end{Bmatrix}, \begin{Bmatrix} d^{dest}_{agr\ meta} \cup d^{dest}_{agr\ vol\ VoIP} \cup d^{dest}_{mng\ 2} \end{Bmatrix}, \begin{Bmatrix} c^{dest}_{storage\ DAS} \cup c^{dest}_{storage\ NAS} \end{Bmatrix}, \begin{Bmatrix} i^{main}_{lc} \cup i^{main}_{mng} \cup i^{main}_{area} \end{Bmatrix} \right\rangle.$$

7. Заключение. Таким образом, практическое использование разработанных моделей для систем мониторинга ИТКС позволяет эффективно комбинировать или агрегировать в одном программно-аппаратном комплексе (ПАК) технологии активного, пассивного и квазипассивного сбора информации, использовать различные БД при обработке результатов мониторинга. Описанные модели легли в осно-

ву модуля квазипассивного мониторинга, входящего в состав распределенной аналитической системы. В сравнении с известными сканерами портов и уязвимостей (acunetics, nmap, zmap, OpenVAS, MaxPatrol, wikto и др.) существенно снижена продолжительность сеанса мониторинга. Отличительной особенностью реализации и комбинированного использования технологий пассивного и квазипассивного мониторинга является использование геобаз на различных этапах (как первичной, так и вторичной) обработки результатов мониторинга, whois-сервисов, централизованное хранение параметров сетевых объектов и типовых профилей в распределенной БД, загрузка/выгрузка дампов мониторинга и результатов обработки в различных форматах.

Все модули ПАК имеют открытый исходный код. В перспективе данный проект планируется развивать в следующих направлениях:

– разработки децентрализованной подсистемы накопления данных об объектах ИТКС;

– разработки метода учета динамики изменения свойств объектов ИТКС;

– разработки технологий многомерного отображения объектов и субъектов ИТКС;

– совершенствования методик анализа сетевого трафика.

Литература

1. *Allen J.M.* OS and Application Fingerprinting Techniques. SANS Institute. 2008.
2. *Arcolano N., Miller B.A.* Statistical Models and Methods for Anomaly Detection in Large Graphs // SIAM Annual Meeting. Minisymposium «Massive Graphs: Big Compute Meets Big Data». 2012.
3. *Beard M.S., Bliss N.T., Miller B.A.* Matched Filtering for Subgraph Detection in Dynamic Networks // Proceedings of the IEEE Statistical Signal Processing Workshop. 2011. pp. 509–512.
4. *Chitpranee R., Fukuda K.* Towards passive DNS software fingerprinting // AINTEC'13 Proceedings of the 9th Asian Internet Engineering Conference. 2013. pp. 9–16.
5. *Gordon L.* Nmap reference guide. URL: http://nmap.org/.
6. P0f v3 (version 3.08b). URL: http://lcamtuf.coredump.cx/p0f3/
7. *Johnson A., Wacek C.* Users Get Routed: Traffic Correlation on Tor by Realistic Adversaries // 20th ACM Conference on Computer and Communications Security. 2013. pp. 337–348.
8. *Kollmann E.* Satori homepage. URL: http://myweb.cableone.net/xnih/.
9. *Kotenko I., Stepashkin M.* Analyzing Vulnerabilities and Measuring Security Level at Design and Exploitation Stages of Computer Network Life Cycle // Lecture Notes in Computer Science. The Third International Workshop «Mathematical Methods, Models and Architectures for Computer Networks Security» (MMM-ACNS-05). Springer-Verlag. vol. 3685. 2005. pp. 317–330.
10. *Matousek P., Rysavy O., Gregr M.* Towards Identification of Operating Systems from the Internet Traffic. IPFIX Monitoring with Fingerprinting and Clustering // Proceedings of the 5th International Conference on Data Communication Networking

(DCNET 2014). Wien: SciTePress – Science and Technology Publications. 2014. pp. 21–27.
11. *Nguyen L.T., Zhang J.* Wi-Fi fingerprinting through active learning using smartphones // Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication. 2013. pp. 969–976.
12. *Ornaghi A.* Ettercap. URL: http://ettercap.github.io/ettercap/.
13. Särökaari N. How to identify malicious HTTP Requests // SANS Institute. 2012. 25 p.
14. *Shaner R.A.* US Patent No. 5991714. 1999.
15. ГОСТ Р 50922-2006. Защита информации. Основные термины и определения // М.: Госстандарт России. 2006.
16. *Дождиков В.Г., Салтан М.И.* Краткий энциклопедический словарь по информационной безопасности // М.: "Энергия". 2012. 240 с.
17. *Комашинский Д.В., Котенко И.В.* Методы интеллектуального анализа данных для выявления вредоносных программных объектов: обзор современных исследований // Вопросы защиты информации. 2013. № 4. С. 21–33.
18. *Смелянский Р.Л.* Модель поведения сетевых объектов в распределённых вычислительных системах // Программирование. 2007. № 4. С. 20–31.
19. *Таненбаум Э., Уэзеролл Д.* Компьютерные сети. Пятое издание // Спб.: "Питер". 2012. 960 с.
20. *Хоар Ч.* Взаимодействующие последовательные процессы // М.: "Мир". 1989. 264 с.
21. *Черемушкин А.В.* Информационная безопасность. Глоссарий // М.: "АВАНГАРД ЦЕНТР". 2013. 322 с.

References

1. Allen J.M. OS and Application Fingerprinting Techniques. SANS Institute. 2008.
2. Arcolano N., Miller B.A. Statistical Models and Methods for Anomaly Detection in Large Graphs. SIAM Annual Meeting. Minisymposium «Massive Graphs: Big Compute Meets Big Data». 2012.
3. Beard M.S., Bliss N.T., Miller B.A. Matched Filtering for Subgraph Detection in Dynamic Networks. Proceedings of the IEEE Statistical Signal Processing Workshop. 2011. pp. 509–512.
4. Chitpranee R., Fukuda K. Towards passive DNS software fingerprinting. AINTEC'13 Proceedings of the 9th Asian Internet Engineering Conference. 2013. pp. 9–16.
5. Gordon L. Nmap reference guide. URL: http://nmap.org/.
6. P0f v3 (version 3.08b). URL: http://lcamtuf.coredump.cx/p0f3/
7. Johnson A., Wacek C. Users Get Routed: Traffic Correlation on Tor by Realistic Adversaries. 20th ACM Conference on Computer and Communications Security. 2013. pp. 337–348.
8. Kollmann E. Satori homepage. URL: http://myweb.cableone.net/xnih/.
9. Kotenko I., Stepashkin M. Analyzing Vulnerabilities and Measuring Security Level at Design and Exploitation Stages of Computer Network Life Cycle. Lecture Notes in Computer Science. The Third International Workshop «Mathematical Methods, Models and Architectures for Computer Networks Security» (MMM-ACNS-05). Springer-Verlag. vol. 3685. 2005. pp. 317–330.
10. Matousek P., Rysavy O., Gregr M. Towards Identification of Operating Systems from the Internet Traffic. IPFIX Monitoring with Fingerprinting and Clustering. Proceedings of the 5th International Conference on Data Communication Networking (DCNET 2014). Wien: SciTePress – Science and Technology Publications. 2014. pp. 21–27.

11. Nguyen L.T., Zhang J. Wi-Fi fingerprinting through active learning using smartphones. Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication. 2013. pp. 969–976.
12. Ornaghi A. Ettercap. URL: http://ettercap.github.io/ettercap/.
13. Särökaari N. How to identify malicious HTTP Requests. SANS Institute. 2012. 25 p.
14. Shaner R.A. US Patent No. 5991714. 1999.
15. GOST R 50922-2006. [Protection of information. Basic terms and definitions]. M.: Gosstandart Rossii. 2006. (In Russ.).
16. Dozhdikov V.G. *Kratkiy entsiklopedicheskiy slovar' po informatsionnoy bezopasnosti* [Short Encyclopedic Dictionary of Information Security]. M.: Energy. 2012. 240 p. (In Russ.).
17. Komashinskiy D.V., Kotenko I.V. [Data mining techniques to identify malicious software Ob-projects: a review of current research]. *Voprosy zashchity informatsii – Problems of information security*. 2013. vol. 4. pp. 21–33. (In Russ.).
18. Smelyanskiy R.L. [Behavioral model of network objects in distributed computing systems]. *Programmirovaniye – Programming.* 2007. vol. 4. pp. 20–31. (In Russ.).
19. Tanenbaum E., Uezeroll D. *Komp'yuternyye seti. Pyatoye izdaniye* [Computer networks. Fifth Edition]. Spb.: Piter. 2012. 960 p. (In Russ.).
20. Hoare C. *Vzaimodeystvuyushchiye posledovatel'nyye protsessy* [Communicating Sequential Processes]. M.: "The World". 1989. 264 p. (In Russ.).
21. Cheremushkin A.V. *Informatsionnaya bezopasnost'. Glossariy.* [Information security. Glossary]. M.: "AVANGUARD CENTER". 2013. 322 p. (In Russ.).

Овчаров Владимир Александрович — к-т техн. наук, докторант кафедры систем сбора и обработки информации, Военно-космическая академия имени А.Ф. Можайского. Область научных интересов: технологии мониторинга сетей, анализ трафика, кластерный анализ, теория вычислительной сложности, расследование инцидентов информационной безопасности. Число научных публикаций — 36. 9823800@inbox.ru; ул. Ждановская, д 13, Санкт-Петербург, 197198; р.т.: +7(812)237-19-60.

Ovcharov Vladimir Aleksandrovich — Ph.D., doctoral student of systems for collecting and processing information department, Mozhaisky Military Space Academy. Research interests: technology network monitoring, cluster analysis, network situational awareness, computational complexity, network forensics, traffic analysis. The number of publications — 36. 9823800@inbox.ru; 13, Zhdanovskay street, St.-Petersburg, 197198, Russia; office phone: +7(812)237-19-60.

РЕФЕРАТ

Овчаров В.А. **Моделирование субъектно-объектного взаимодействия в сетевых инфраструктурах.**

В работе рассматривается задача разработки моделей поведения объектов сетевых инфраструктур (пользователей, устройств и ресурсов) по результатам анализа субъектно-объектного взаимодействия. В качестве решения данной задачи в части идентификации типов сетевых объектов и операций взаимодействия предлагается графовая модель поведения, в части деанонимизации отношений взаимодействующих объектов предложены предикатные модели состояний объектов информационно-телекоммуникационной сети (ИТКС) на основе отношений между экземплярами.

На основе анализа решаемых задач, архитектуры, особенностей программной реализации и специфики взаимодействия функциональных подсистем современных систем мониторинга ИТКС предложена таксономия процессов сбора, обработки и хранения данных мониторинга, позволяющая синтезировать структурные логико-графические модели систем мониторинга ИТКС различной архитектуры и назначения.

SUMMARY

Ovcharov V.A. **Simulation of Subject-Object Interaction in Network Infrastructures.**

The problem of modeling the behavior of objects network infrastructures (users, devices, and resources) for the analysis of subject-object interaction is considered. As a solution to this problem in terms of identifying the types of network facilities and operations interaction graph model of behavior is proposed, in terms of relations disclosure of anonymity interacting objects predicate state model object information and telecommunications network (ITN) on the basis of relations between instances is offered.

Based on the analysis of tasks, architecture, features software implementation and specific interaction of functional subsystems of modern monitoring systems of ITN taxonomy of the collection, processing and storage of monitoring data, which allows to synthesize structural logic-graphic model of monitoring systems ITN different architecture and destination, is proposed.

УДК 316.77:519.7

С.В. Пилькевич, М.А. Еремеев

МОДЕЛЬ СОЦИАЛЬНО ЗНАЧИМЫХ ИНТЕРНЕТ-РЕСУРСОВ

Пилькевич С.В., Еремеев М.А. **Модель социально значимых Интернет-ресурсов.**
Аннотация. Рассматривается модель социально значимых Интернет-ресурсов. Модель предназначена для исследования процессов коммуникативного и когнитивного взаимодействия пользователей современных социально значимых Интернет-ресурсов. Комбинирование результатов, полученных на базе теорий когнитивного соответствия в социальной психологии, исследований восприятия и забывания информации в физиологии, а также основных законов теории информации позволило разработать модель, в равной степени адекватно описывающую информационно-психологическое взаимодействие участников различных видов социальных ресурсов: форумов, социальных сетей, блогов и Интернет СМИ.
Ключевые слова: социальная сеть, социально значимый интернет-ресурс, коммуникационное взаимодействие, когниция, социальная психология.

Pilkevich S.V., Eremeev M.A. **Model of Socially Important Internet Resources.**
Abstract. This paper represents a model of socially important Internet resources. The model is designed to study the processes of communicative and cognitive users interact modern socially important Internet resources. Combining the results obtained on the basis of cognitive theories of conformity in social psychology, studies of perception and forgetting information in physiology as well as basic laws of information theory has allowed us to develop the model, equally adequately describe the information-psychological interaction of participants in various types of social resources: forums, social networks, blogs and online media.
Keywords: social network, socially important Internet resource, communication, cognize, social psychology.

1. Введение. В настоящее время социальный мир человека существенным образом изменился, современная ситуация такова, что значимость субъекта определяется его информированностью и принадлежностью к виртуальным сообществам, которые становятся объективной реальностью и оказывают влияние на системные свойства социума.

Современные системы связи и телекоммуникаций, предоставляя абонентам широчайшие возможности по оперативному обмену большими объемами информации, задают основной вектор их социального развития и поведения, интенсивности проявления конформизма, изменения мнений по важнейшим социальным, экономическим и духовным вопросам.

Если еще совсем недавно уровень развития государства определялся количеством природных ресурсов и возможностями полноценного их использования, затем уровнем владения передовыми производственными технологиями, то теперь все чаще во главу угла ставятся создание и владение информацией, новейшие способы ее обработки, разработка новых парадигм и технологий.

В этой ситуации становится актуальным изучение феномена информации и свойств информационных потоков не с точки зрения computer science, а с точки зрения междисциплинарного подхода. Другими словами изучение структуры и свойств процесса производства информации, взаимодействия информационных потоков с социальной средой, адаптации человека к жизни в таком «информатизированном» обществе [1].

Понимание механизмов влияния на общество, его структурирование, возможности управления как отдельными социальными группами, так и обществом в целом, становится одной из важнейших задач ближайшего будущего, а создание базы для развития этого направления является актуальнейшей задачей уже сейчас.

При моделировании информационных потоков в социальных средах необходимо основываться на результатах, в первую очередь, математики, психологии и социологии, а также экономики. Это обусловлено следующими основными причинами:

– деятельность, связанная с производством, распространением, обработкой информации стала важнейшим сектором экономики;

– информационное пространство становится основной ареной соперничества государств, элит, транснациональных корпораций, наравне с сушей, морской поверхностью и его глубинами, воздушным пространством и космосом;

– особое внимание привлекает информационное управление обществом или отдельными социальными группами. Эра персональных компьютеров и глобальных телекоммуникационных сетей, вероятно, несет новые глобальные перемены [1].

Выше изложенное позволяет сделать вывод о необходимости разработки модели средств межличностной коммуникации с позиций междисциплинарного подхода применительно к современным социальным сетям.

2. Терминология и классификация. Понятие «социальная сеть» (применительно к реализации сетевых сервисов) появилось сравнительно недавно, в середине 90-х годов прошлого века. Социальная сеть - это платформа, онлайн-сервис или веб-сайт, предназначенные для построения, отражения и организации социальных взаимоотношений, визуализацией которых являются социальные графы [2].

Родственным термином является понятие «социальные медиа». Под социальными медиа понимается вид массовой коммуникации, осуществляемый посредством Интернета и предоставляющий возможность публикации, обмена и обсуждения контента широким кругом пользователей.

На практике, технологии, лежащие в основе реализации социальных сетей и социальных медиа, тесно переплетены, что приводит к появлению и широкому распространению мультисервисных социально значимых Интернет-ресурсов, к которым можно отнести: блоги, онлайновые социальные сети, хостинги, форумы и др.

Несмотря на ряд общих черт, средства интернет-коммуникации, входящие в категорию социально значимых Интернет-ресурсов, могут значительно различаться по функциям и возможностям, которые они предоставляют пользователям, а также принципам и нормам, действующим на каждом конкретном сайте. В связи с этим целесообразно говорить о системе классификации социально значимых Интернет-ресурсов по нескольким основаниям (по степени анонимности, по принципам предоставления доступа к контенту, по функциональным возможностям сервисов), схема [3] которой представлена на рисунке 1.

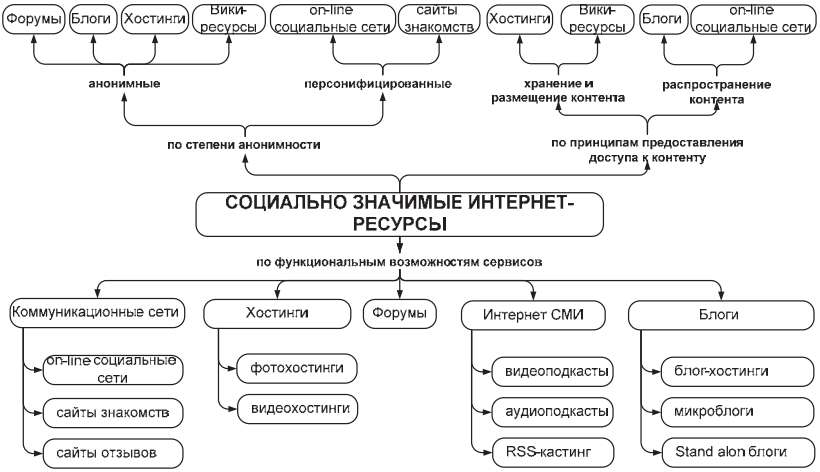


Рис. 1. Классификация социально значимых Интернет-ресурсов

К наиболее весомым игрокам на рынке социально значимых Интернет-ресурсов, безусловно, необходимо отнести: из зарубежных - Facebook, Twitter, Google+ и LinkeIn, из отечественных – Вконтакте и Одноклассники. Количество пользователей перечисленных ресурсов измеряется миллионами.

Решая задачи коммуникации, современные социально значимые Интернет-ресурсы обладают богатым набором средств публикации и управления контентом. К основным средствам следует отнести: распространения, оформления и публикации информации, поддержания обратной связи, возможность формирования и изменения списка других поль-

зователей, а также персонификации автора сообщения.

Таким образом, современные социально значимые Интернет-ресурсы предоставляют пользователю все необходимые инструменты и средства для публикации в Интернете информации любого типа (как текстовой, так и мультимедийной), связывания нового материала с уже опубликованной информацией (причем опубликованной не обязательно самим пользователем) посредством гиперссылок, а также объединения нескольких типов информации в одной публикации.

3. Модель пользовательского аккаунта. Персонификация сообщений достигается за счет регистрации личного профиля (аккаунта) пользователя, представляющего собой одну из характерных особенностей социально значимых Интернет-ресурсов.

Процедура регистрации пользователя сопряжена с необходимостью заполнения своеобразной персональной анкеты, при этом информация, заносимая пользователем подразделяются на обязательную и дополнительную. Как правило перечень полей обязательных для заполнения является идентичным на различных социально значимых Интернет-ресурсах [3] (см. таблицу 1).

Таблица 1. Персонифицированная информация, необходимая для регистрации аккаунта

Требования при регистрации		
обязательные	дополнительные	
Имя, Фамилия *(Name, Surname)*	Место работы, должность *(Work)*	Адрес, местоположение *(Geo)*
Дата рождения *(Bday)*	Специализация и проф. навыки *(Prof)*	Контакты, друзья *(Comm), (Friend)*
Адрес эл. почты *(Email)*	Навыки, знания, цели *(Skill)*	Присутствие на других сайтах *(Web)*
Номер мобильного телефона *(Phone)*	Отрасли в которых вы наиболее компетентны *(Ind)*	Выбрать интересы (предлагаются варианты) *(Hobby)*
Город *(City)*	Образование, где учились, год *(Course)*	Членство в обществах и ассоциациях *(Assoc)*
Школа, университет *(Sch), (Un)*	Служба в армии *(Mil)*	Брак и семья, дети, национальность *(Family), (Eth)*

На основе обобщения информации о параметрах регистрации и заполнения пользовательского профиля в более чем 140 социально

значимых Интернет-ресурсах [3] предложена следующая формальная модель пользовательского аккаунта:

$$A=<Pers, Cont, CoC, Prof>,$$

где *Pers* =<*Name, Surname, Bday*> - подмножество идентифицирующей информации;
Cont =<*Email, Phone, City, Geo, Comm, Web*> - подмножество контактной информации;
CoC =<*Friend, Family*> - подмножество информации о социальных связях;
Prof =<*Sch, Un, Work, Prof, Skill, Ind, Course, Mil, Hobby, Assoc, Eth*> - подмножество информации об уровне образования, профессиональных компетенциях и предпочтениях,
где $Email = \{Email_i\}$; $Phone = \{Phone_j\}$; $Skill = \{Skill_k\}$; $Ind = \{Ind_l\}$; $Course = \{<Course_1, Y_1>,\ldots, <Course_n, Y_n>\}$; $Friend=\{<Name_1, Surname_1>,\ldots, <Name_m, Surname_m>\}$; $Web=\{Web_p\}$; $Assoc=\{Assoc_r\}$; $Family=\{<Family_1, Role_1>,\ldots, <Family_u, Role_f>\}$; $Sch=\{<Sch_1, Y_1>,\ldots, <Sch_s, Y_s>\}$; $Un=\{<Un_1, Y_1>,\ldots, <Un_t, Y_t>\}$, $Y \in [1900, 2015]$;

$Role \in \{муж, жена, дочь, сын, отец, мать\}$, $i, j, k, l, m, n, p, r, u, f, s, t \in \mathbb{N}$.

Необходимо отметить, что параметры, входящие составными элементами в модель пользовательского аккаунта в ряде случаев имеют функциональные зависимости друг с другом см., например, [4].

4. Анализ коммуникационных возможностей социально значимых Интернет-ресурсов. Модель распространяемой информации. Одной из характерных особенностей социально значимых Интернет-ресурсов является предоставление пользователям практически полного спектра возможностей для обмена информацией (размещение фотографий, видео- и текстовых записей, организация тематических сообществ, обмен личными сообщениями и т.п.). Исследования [3] показывают, что наиболее распространенными параметрами социальных сервисов в части, касающейся публикации и управления контентом, являются:

- публикация текстовой и графической информации;
- возможность комментировать и (или) высказывать свое мнение по отношению к опубликованной информации;
- поддержка системы тегов;
- формирование тематических групп пользователей;
- публикация личных сообщений.

При этом представляет интерес классификация информации, формируемой пользователями сети Графическое представление данной классификации изображено на рисунке 2.

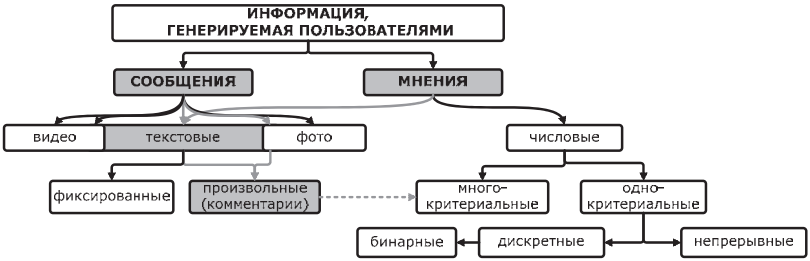


Рис. 2. Виды информации, генерируемой пользователями социально значимых Интернет-ресурсов

Исходя из выше изложенного, видится целесообразным формализовать модель распространяемой в социально значимых Интернет-ресурсах информации в следующем виде:

$$I=<Pers, M, T>,$$

где *Pers* - подмножество идентифицирующей ***A*** информации;
A – актор сети (аккаунт пользователя);
$M \in \{M_T, M_F, M_V, M_T \cup M_F, M_T \cup M_V\}$ - мнения / комментарии (текст, фото, видео и т.д.);
T – отметка времени;
$K_i=f_A(M)$ – когниция, $i \in \mathbb{N} \cup 0$.

Множество мнений/комментариев, представлено главным образом, текстовой информацией, а также фотографиями и видео роликами, имеющими краткие текстовые пояснения. Отметим, что среди современных социальных Интернет-ресурсов обмен мультимедийным контентом поддерживает не более 40% [3].

Предлагаемая модель носит универсальный характер и не является локализованной версией пригодной для какой-то конкретной языковой культуры. Информация, публикуемая пользователями, является, по сути, отражением их концептуальной картины мира, которая получается в результате прямого познания окружающей действительности.

Исходя из [5] концептуальная картина мира:

1) определяется как когнитивная,

2) складывается в сознании индивидуума под воздействием интеракции речевого взаимодействия,

3) представляет собой результат когниции (познания) действительности,

4) выступает в виде совокупности упорядоченных знаний – концептосферы.

Под термином когниция понимается, элемент знания (данные, усвоенные сознанием) [6]. Подходы к разработке алгоритмов, реализующих функциональное отображение f_A достаточно разнообразны [7–13].

В случае использования системы тегов формализация процедуры получения когниций из текстовых сообщений существенно упрощается [14, 15]. В таблице 2 представлен пример тегов системы микроблогов Twitter [13].

Таблица 2. Теги системы микроблогов Twitter и примеры их использования

Вид тега	@-ссылки	Ретвиты	Слэштеги	Хэштеги
Пример использования	*Some text. /via @User*	*RT @User Some text*	/via /by /cc (/for) /thx /ht (/hat tip)	*#web20* *#haiku* *#haiti*

Исходя из вышеизложенного, приходим к необходимости связи моделей аккаунтов и информации, генерируемой пользователями социально значимых Интернет-ресурсов. Отметим, что модель аккаунтов содержит информацию о коммуникационном взаимодействии пользователей $[\alpha_{ik}]$.

Пусть A_i, A_k – аккаунты пользователей сети (элементы социума), тогда α_{ik} - социальная связь между элементами A_i и A_k, такая, что

при $\alpha_{ik} < 0$ - отрицательное,

при $\alpha_{ik} = 0$ - нейтральное,

при $\alpha_{ik} > 0$ - положительное отношение.

$\alpha_{ik} = F(Friend_i, Friend_i, Family_k, Family_k, f_i(I_j))$, где I_j - сообщение/комментарий, содержащий проблемное утверждение (тезис /когницию K_j). Кроме того, по аналогии с предложенным в [16] коэффициентом, устанавливающим толщину стрелки, связывающей двух индивидов – объекты социометрического наблюдения социума, определим статистическую величину интенсивности отношений, существующих между двумя индивидуумами A_i и A_k следующим образом:

$$a_{ik} = N \log_2 \frac{N_{ik}}{N_0},$$

где N_{ik} - величина сообщения, или количество переданных знаков, N_0 - эталонная величина, определяемая обычно как среднее значение ин-

тенсивности связей внутри группы при данных условиях (в течение временного диапазона мониторинга) [16]. Абсолютное значение константы *N* различно для различных индивидуумов и зависит от условий, в которых находится испытуемый, аналогично константе K закона Вебера-Фехнера [17].

Модель информации содержит сведения о семантике циркулирующих в Интернет-ресурах сообщений (мнений, комментариев и т.д.).

Представляется оправданным рассмотреть прагматику как аспект взаимодействия пользователей (моделируемых их аккаунтами) и сообщений (модель информации, генерируемой пользователями).

Таким образом, требуется построить модель информационно-психологического отношения *B*, такого, что $B \subseteq A \times K$, где $A = \{A_1, A_2, ..., A_l\}, l \in \mathbb{N}$, $K = \{K_1, K_2, ..., K_k\}, k \in \mathbb{N}$ и, как следствие, рассмотреть когнитивную сферу взаимодействия пользователей социально-значимых Интернет-ресурсов.

Рассмотрим предложенную схему взаимодействия пользователей и информации в виде графа с нагруженными дугами, отображающими когнитивные и коммуникативные отношения, которые могут быть реальными или виртуальными (предполагаемыми) (см. рисунок 3).

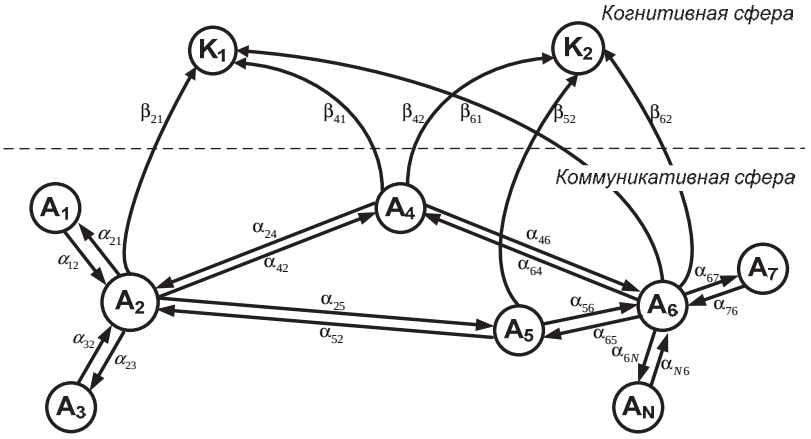


Рис. 3. Схема системы психологических отношений в социуме (социально значимых Интернет-ресурсах)

Предложенная схема была построена с учетом того, что когнитивное моделирование в концептуальном плане базируется на теориях когнитивного соответствия в социальной психологии: теории когнитивного диссонанса Л. Фестингера [18], теории структурного баланса Ф. Хайдера [19], теории коммуникационных актов Т. Ньюкома [20], теории конгруэнтности Ч. Осгуда и П. Танненбаума [21, 22]. Базовыми постулатами для которых являются следующие:

1) человек обладает способностью к восприятию, усвоению и переработке информации;

2) человек всегда стремится к психическому равновесию, т.е. к достижению внутренней связности, логичности, непротиворечивости своей картины мира;

3) когнитивные элементы (знания) не всегда органично соответствуют личностной картине мира, что вызывает противоречие между ними (диссонанс) и напряженность, требующую разрешения, которое осуществляется в форме побуждения к некоторым действиям – поведению [6].

5. Моделирование динамики психологической ситуации. Всякое взаимодействие между пользователями сети может быть представлено как элементарная информационно-психологическая акция (ИПА). Участниками элементарной ИПА являются два пользователя (социальных элемента), один из которых условно обозначен как *S*-субъект, а другой как *O*-объект ИПА.

Пусть существует некоторое элементарное утверждение K из когнитивного алфавита проблемной ситуации, которое является контекстом ИПА.

В течение предыдущих взаимодействий между элементами сложились некоторые отношения α_{OS} и α_{SO}, а также сформировались личные отношения к упомянутому утверждению - β_{OK} и β_{SK}.

Описанная ситуация в виде элементарного графа-триады с нагруженными дугами представлена на рисунке 4.

В некоторый момент времени регистрируется элементарная ИПА. Под элементарной ИПА понимается такое принудительное изменение информационно-психологического баланса, в котором участвуют *S*, *O* и воспроизводится одно элементарное утверждение K, сопровождаемое однозначно трактуемым отношением согласия (одобрения) или несогласия (неодобрения) [6].

Акция представляет собой доступное для восприятия *O* целенаправленное проявление *S* некоторого утверждения - *K* и выражения оценочного отношения к этому знанию - β_{SK} (см. рисунок 5).

В результате ИПА происходит изменение информационно-психологического баланса в триаде, т.е. изменяются исходные значения α_{OS} и α_{SO}, β_{OK} и β_{SK} (см. рисунок 6).

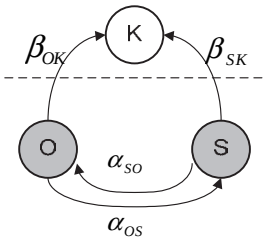


Рис. 4. Исходная схема отношений в триаде O-S-K (до ИПА)

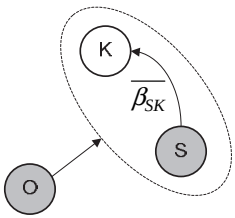


Рис. 5. Элементарная ИПА

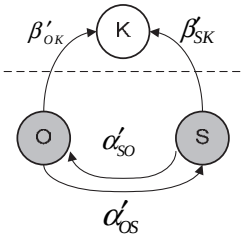


Рис. 6. Отношения в триаде O-S-K после ИПА

Данные теории когнитивистского направления [17-21] постулируют существование и естественное поддержание информационно-псиологического баланса в социальных системах и позволяют формализовать механизм изменения значений α_{OS}, α_{SO}, β_{OK} и β_{SK}.

Вычисление балансирующих изменений параметров когнитивных и межэлементных отношений, осуществляется на основе дифференциалов Осгуда и применительно к предложенной модели представлено в [7] следующим образом:

$$\Delta_{OK} = \frac{|\alpha_{OS}|}{|\beta_{OK}|+|\alpha_{OS}|} \times |\beta_{OK} - \alpha_{OS}|, \quad \Delta_{OS} = \frac{|\beta_{OK}|}{|\beta_{OK}|+|\alpha_{OS}|} \times |\beta_{OK} - \alpha_{OS}|,$$

где Δ_{OK} и Δ_{OS} - социометрические приращения отношений объекта к субъекту воздействия и к утверждению, содержащемуся в материале ИПА.

Рассмотрение всех вариантов триад (таблица 3), приводимых в теории структурного баланса Ф.Хайдера и направлений их изменения в соответствии с теорией коммуникативных актов Т.Ньюкомба позволяет получить формулы для вычисления итоговых значений α'_{OS} и β'_{OK}.

Таблица 3. Варианты триад и направления их изменения

№	Описание триады	Формулы вычисления принудительной коррекции отношений
	Сбалансированные исходные триады	
1.	Объект изначально позитивно относится к субъекту воздействия и к утверждению в материале ИПА, субъект выражает позитивное отношение к утверждению в материале ИПА.	У объекта усиливаются позитивные тенденции отношения как к субъекту, так и к утверждению: $\alpha'_{OS} = \alpha_{OS} + \Delta_{OS}; \beta'_{OK} = \beta_{OK} + \Delta_{OK}$
2.	Объект изначально позитивно относится к субъекту воздействия и негативно к утверждению в материале ИПА, субъект также выражает негативное отношение к утверждению в материале ИПА.	У объекта усиливаются позитивные тенденции отношения к субъекту и негативные к утверждению: $\alpha'_{OS} = \alpha_{OS} + \Delta_{OS}; \beta'_{OK} = \beta_{OK} - \Delta_{OK}$
3.	Объект изначально негативно относится к субъекту воздействия и негативно к утверждению в материале ИПА, субъект же выражает позитивное отношение к утверждению в материале ИПА.	У объекта усиливаются негативные тенденции отношения к субъекту и негативные к утверждению: $\alpha'_{OS} = \alpha_{OS} - \Delta_{OS}; \beta'_{OK} = \beta_{OK} - \Delta_{OK}$
4.	Объект изначально негативно относится к субъекту воздействия и позитивно к утверждению в материале ИПА, субъект же выражает негативное отношение к утверждению в материале ИПА.	У объекта усиливаются негативные тенденции отношения к субъекту и позитивные к утверждению: $\alpha'_{OS} = \alpha_{OS} - \Delta_{OS}; \beta'_{OK} = \beta_{OK} + \Delta_{OK}$
	Несбалансированные исходные триады	
5.	Объект изначально позитивно относится к субъекту воздействия и негативно к утверждению в материале ИПА, но субъект выражает позитивное отношение к утверждению в материале ИПА.	У объекта наблюдается тенденция ухудшения отношения к субъекту и тенденция улучшения отношения к утверждению: $\alpha'_{OS} = \alpha_{OS} - \Delta_{OS}; \beta'_{OK} = \beta_{OK} + \Delta_{OK}$
6.	Объект изначально позитивно относится к субъекту воздействия и позитивно к утверждению в материале ИПА, но субъект выражает негативное отношение к утверждению в материале ИПА.	У объекта наблюдается тенденция ухудшения отношения к субъекту и тенденция ухудшения отношения к утверждению: $\alpha'_{OS} = \alpha_{OS} - \Delta_{OS}; \beta'_{OK} = \beta_{OK} - \Delta_{OK}$
7.	Объект изначально негативно относится к субъекту воздействия и позитивно к утверждению в материале ИПА, но субъект также выражает позитивное отношение к утверждению в материале ИПА.	У объекта наблюдается тенденция улучшения отношения к субъекту и тенденция ухудшения отношения к утверждению: $\alpha'_{OS} = \alpha_{OS} + \Delta_{OS}; \beta'_{OK} = \beta_{OK} - \Delta_{OK}$
8.	Объект изначально негативно относится к субъекту и к утверждению в материале ИПА, субъект также выражает негативное отношение к утверждению в материале ИПА.	У объекта наблюдается тенденция ослабления негативного отношения к субъекту и к утверждению: $\alpha'_{OS} = \alpha_{OS} + \Delta_{OS}; \beta'_{OK} = \beta_{OK} + \Delta_{OK}$

6. Социальная диффузия как «когнитивный шум». Социальный элемент, восприняв некоторую когницию, становится не только носителем сформированного к данной когниции психоэмоционального отношения, постепенно утрачивающего актуальность (кривая Эббингауза [23]), но и является возможным источником вторичного распространения этого утверждения и своего текущего отношения к нему.

При достижении элементом определенного уровня психоэмоционального отношения к когниции элемент «резонирует», становясь вторичным источником ИПА, которая не всегда тождественна первичной. Порог «резонирования» социального элемента - параметр, вычисляемый на основании непрерывно накапливаемых статистических данных [6].

Социальная коммуникативная активность может быть охарактеризована регистрируемой частотой вступления в коммуникативные отношения с окружающими социальными элементами по инициативе субъекта. Чем более активен социальный элемент, тем чаще он становится участником коммуникативных актов и тем чаще периодически воспроизводит то или иное утверждение и свое отношение к нему.

Если впоследствии не происходит принудительной репродукции, то можно утверждать, что чем больше времени проходит с момента генерации психоэмоционального «пика» отношения к утверждению, тем сильнее снижается частота его вторичных воспроизведений, т.е. периоды «молчания» возрастают.

Чем ближе к моменту синтеза отношения, тем частота инициируемых коммуникативных взаимодействий по поводу этой когниции выше.

Функция активизации памяти и воспроизведения утверждения может быть представлена в виде некоторой псевдогармоники с нарастающим периодом.

На основании выше изложенного в [6] приводится функция, описывающая когнитивный «резонанс», имеющая следующее выражение:

$$\beta_{ij}(t) = \beta_{ij}(t_0)\cos\left(\rho_i\sqrt{t-t_0}\right)e^{-\lambda_i(t-t_0)},$$

где t_o - время последнего принудительного изменения информационно-психологического баланса элемента A_i по поводу утверждения K_j, вызвавшего когнитивный резонанс; t - время на которое осуществляется контрольный анализ ситуации; $\lambda_i(t)$ - коэффициент «старения и

утраты актуальности утверждения K_j »; $\rho_i(t)$ - частотный коэффициент когнитивного резонанса A_i . Графическое представление функции $\beta_{ij}(t)$, ограниченной кривыми забывания (пунктир) приведено на рисунке 7.

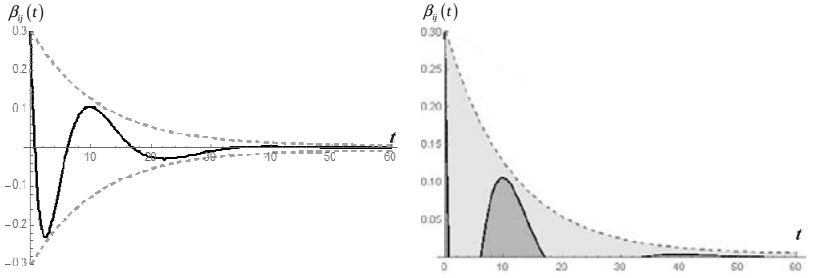


Рис. 7. Предполагаемый уровень выражаемого A_i отношения $\beta_{ij}(t)$ к утверждению K_j

Положительные полупериоды на графике соответствуют отрезкам времени, в течение которых возможны проявления коммуникативной активности элемента A_i по поводу утверждения K_j .

Отрицательные полупериоды моделируют такие отрезки времени, в течение которых проявление когнитивной активности маловероятно.

Учитывая вышесказанное, функция, моделирующая когнитивную активность, приобретает следующий вид:

$$\beta_{ij}(t) = \tfrac{1}{2}\beta_{ij}(t_0)\left(\left|\cos\left(\rho_i\sqrt{t-t_0}\right)\right| + \cos\left(\rho_i\sqrt{t-t_0}\right)\right)e^{-\lambda_i(t-t_0)}.$$

Для всего социума модель предполагаемой диффузной активности элементов может быть представлена в графическом виде (см. рисунок 8).

Отметим, что при $\rho = 1$ функция когнитивной активности представляет собой монотонно убывающую кривую, стремящуюся к $\beta_{ij}(t) = 0$. Увеличение значения коэффициента ρ приводит к появлению псевдогармоники с нарастающим периодом, при этом временной интервал $t - t_0$, характеризующий длительность когнтивного резонанса с момента начала наблюдения (t_0) до полного его затухания соответ-

ствует периоду «диффузии инноваций». Различные значения коэффициента ρ могут интерпретироваться как различные пользователи A_i.

Таким образом, представленный на рисунке 8, график диффузной активности элементов визуализирует динамику активности 12

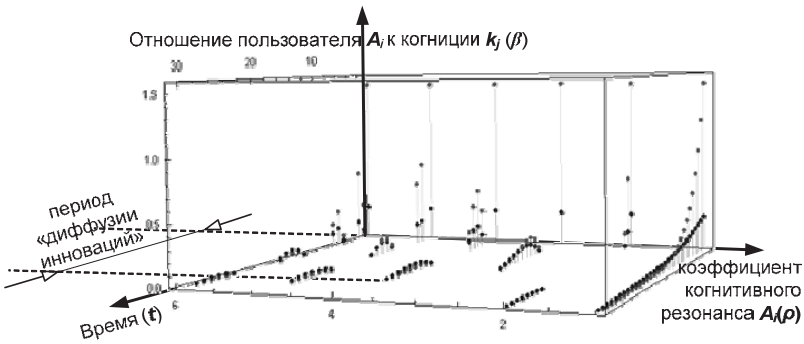


Рис. 8. Графическая интерпретация функции, моделирующей когнитивную активность (информационно-психологическое отношение *B*)

пользователей $\{A_i\}_1^{12}$, разбитых на 6 пар, характеризующихся значениям коэффициента $\rho \in \{1,2,3,4,5,6\}$ и начальными значениями личных отношений к утверждению $\beta_{2i+1,j}(t_0) = 0.3$ и $\beta_{2i,j}(t_0) = 1.5$. Причем указанные начальные значения пользователей синхронизированы. Значения коэффициентов $\lambda_i(t)$ также отличаются: $\lambda_{2i+1}(t) = 0.1$, $\lambda_{2i}(t) = 0.25$.

Функция $\beta_{ij}(t)$, моделирующая когнитивную активность пользователей зависит от трех переменных (коэффициентов ρ и λ, времени *t*), соответственно в трехмерном пространстве может быть представлена только соответствующими сечениями. Представляется, что на рисунке 8 функциональные зависимости, связанные с коэффициентом λ отображены в наименьшей степени. Восполним данный недостаток. График функции $\beta_{ij}(t)$ при фиксированных значениях $\beta_{ij}(t_0) = 0.3$ и $\rho = 3$, представлен на рисунке 9.

Исследование графика функции, изображенного на рисунке 9, позволяет сделать вывод о том, что представленные на нем монотонно убывающие области образованы семействами кривых Эббингауза и характеризуют процессы забывания информации (с течением времени

кривые становятся все менее пологими, что говорит об увеличении скорости забывания информации).

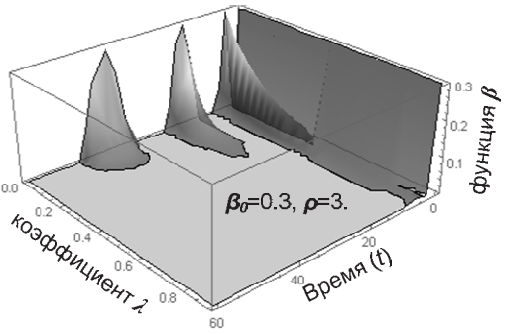
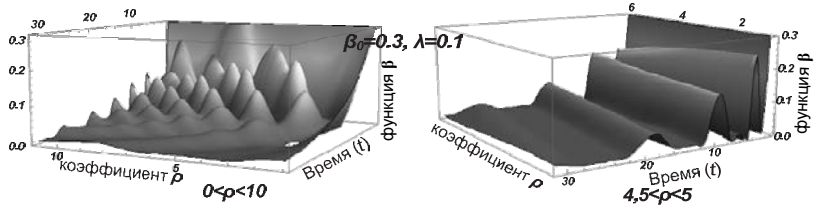


Рис 9. Сечение функции когнитивной активности

С учетом выше изложенного, график диффузной активности группы пользователей (см. рисунок 10 а) не позволяет использовать его в практической деятельности.

Тем не менее, учитывая, что индивидуальные коэффициенты «старения и утраты актуальности утверждения», а также частотные коэффициенты когнитивного резонанса функционально связаны с уровнем образования, профессиональными компетенциями и предпочтениями пользователей, то представляется оправданным для однородных групп существенным образом сужать диапазоны упомянутых коэффициентов. Результат учета особенностей целевой аудитории представлен на рисунке 10 б.



а) б)

Рис. 10. Графическая модель диффузной активности однородной группы пользователей

Таким образом, учет элементов *Sch, Un, Work, Prof, Skill, Ind, Course, Mil, Hobby, Assoc, Eth* модели пользователей A_i позволяет осуществлять учет фонового значения $\beta_{ij}^{\Phi}(t)$ при проведении модели-

рования процессов формирования общественного мнения и решении рядя смежных задач.

В результате, социальная диффузия представляется как некоторый «когнитивный шум», оказывающий влияние на сознание каждого из элементов социума в форме интегральной редукции. Внутрисоциумный «информационный шум» стимулирует фактор социального конформизма или естественное для человеческих сообществ явление редукции сознания [24]. Конформизм, определяемый как изменение поведения или убеждения в результате реального или воображаемого давления группы, подтверждает справедливость гипотезы об уменьшении диапазонов коэффициентов ρ и λ.

7. Заключение. Представленная модель социально значимых Интернет-ресурсов в целом согласуется с классификацией, приведенной на рисунке 1.

В равной степени адекватно описываются информационно-психологическое взаимодействие участников следующих видов ресурсов:

– форумов – модель аккаунтов сводится к множеству акторов сети: $\alpha_{ik}=0$, $A=<Pers>$, т.е. $Cont = \varnothing, CoC = \varnothing, Pof = \varnothing$, а также $|K_l| \approx |M_l|$;

– социальных сетей общения – модель информационно-психологического отношения позволяет исследовать вопросы формирования групп, коалиций, обсуждений тем на стенах и проч.: $\alpha_{ik} \neq 0$, $A=<Pers>$, т.е. $Cont = \varnothing, CoC = \varnothing, Pof = \varnothing$, а также $|K_l| \approx |M_l|$;

– блогов и Интернет СМИ – преобладающую роль играет модель сообщения: $\beta_{ij} = 0$, $|K_l| << |M_l|$.

Остается открытым вопрос о целесообразности введения дополнительных характеристик в модель аккаунта для корректного описания «модели привратника», характерной для моделирования СМИ и хостингов (редакторы, модераторы и администраторы социальных сервисов), а также лидеров мнений, фигурирующих в моделях диффузии инноваций [25].

В частности в [26, 27] делается вывод о том, что в большинстве случаев лидеры мнений лишь умеренно «важнее» обычных пользователей (за исключением некоторых исключительных случаев).

Если этот аспект сети принципиально важен, то метка соответствующего узла может быть получена на основе анализа топологии коммуникационной сети пользователей, в противном случае игнорируется.

Литература

1. *Митин Н.А.* Новые модели математической психологии и информационные процессы // М.: ИПМ РАН. 2013.
2. Социальная сеть. URL: https://ru.wikipedia.org/ (дата обращения: 12.01.2015).
3. Отчет о НИР «Исследование и обоснование принципов корпоративной этики в социальных сетях для военнослужащих и гражданского персонала ВС РФ». Ч.1. // СПб.: ВКА им. А.Ф. Можайского. 2014. 165 с.
4. *Коршунов А.В.* Задачи и методы определения атрибутов пользователей соц. сетей // Труды 15-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2013). Ярославль: ЯГТУ. 2013. С. 183–193. URL: http://ceur-ws.org/Vol-1108/paper23.pdf (дата обращения: 13.10.2014).
5. *Огнева Е.А.* Когнитивное моделирование концептосферы художественного текста: 2-е изд. дополн. // М.: Эдитус. 2013. 282 с.
6. *Семашко К.В., Шеремет И.А.* Математическое моделирование информационно-психологических отношений в социума // М.: Наука. 2007. 157 с.
7. ABBYY Intelligent Search SDK. URL: http://www.abbyy.ru/isearch/compreno/ (дата обращения: 23.12.2014).
8. ABBYY Compreno. URL: http://www.dialog-21.ru/digests/dialog2012/materials/pdf/anisimovich.pdf (дата обращения: 23.12.2014).
9. ABBYY Intelligent Tagger SDK. URL: http://www.abbyy.ru/adx/aspx/adxgetmedia.aspx?DocID=f6bad99e-d66a-4da0-9112-4c6fc15e1f72 (дата обращения: 24.12.2014).
10. *Анисимович К.В., Дружкин К.Ю., Зуев К.А., Минлос Ф.Р., Петрова М.А., Селегей В.П.* Синтаксический и семантический парсер, основанный на лингвистических технологиях // Международная конференция по компьютерной лингвистике «Диалог». URL: http://www.dialog-21.ru/digests/dialog2012/materials/pdf/Anisimovich.pdf (дата обращения: 24.12.2014).
11. *Ермакова Л.М.* Методы извлечения информации из текста // Вестник Пермского университета. 2012. Вып.1(9). С. 77-84. URL: http://dspace.nsu.ru:8080/jspui/bitstream/nsu/202/1/278_84017.pdf (дата обращения: 29.12.2014).
12. *Леонова Ю.В., Федотов А.М.* Извлечение знаний и фактов из текстов диссертаций и авторефератов для изучения связей научных сообществ // Труды 15-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2013). Ярославль: ЯГТУ. 2013. С. 32–41. URL: http://ceur-ws.org/Vol-1108/paper5.pdf (дата обращения: 13.10.2014).
13. *Лукашевич. Н.В.* Извлечение знаний и фактов из текстов // Научно-исследовательский вычислительный центр МГУ им. М.В.Ломоносова. URL: http://www.slideshare.net/msucsai/2007-12 (дата обращения: 13.10.2014).
14. *Атягина А.П.* Твиттер как новая дискурсивная практика в сети интернет // Вестник Омского университета. 2012. №4. С. 203–208.
15. *Коршунов А.В.* Извлечение ключевых терминов из сообщений микроблогов с помощью Википедии // Труды Института системного программирования РАН. 2011. № 20. С. 269–282.
16. *Моль А.* Социодинамика культуры // М.: Издательство «Прогресс». 1973. 405 с.
17. *Головин С.Ю.* Словарь практического психолога // Минск: Харвест. 1998. 800 с.
18. *Festinger L.* Theory of Cognitive Dissonance // Standford. CA. Standford University Press. 1957.
19. *Heider F.* The psychology of interpersonal relations // N.Y. 1958.

20. *Newcomb T.M.* An approach to the study of communicative acts // Psychological Review 1953. vol. 60. pp. 393–404.
21. *Osgood C.E., Tannenbaum P.* The priciple of congruity in the prediction of attitude change // Psychological review. 1955. vol. 62. pp. 42-55.
22. *Osgood C.E., Suci G. Tannenbaum P.* The measurement of meaning // Chicago. «Semantic differential technique». 1957. 342 p.
23. *Бартлетт Ф.* Человек запоминает // Хрестоматия по психологии памяти // М.: Из-во МГУ. 2002. С. 292–303.
24. *Ермаков Ю.А.* Манипуляция личностью: смысл, приемы, последствия // Екатеринбург: Издательство Уральского университета. 1995. 136 с.
25. *Пилькевич С.В., Ломако А.Г.* Модели коммуникации при обеспечении защиты от негативного информационного воздействия // Методы обеспечения информационной кибербезопасности: (дополнительный выпуск). М.: КомКнига. 2013. Т. 27. С. 475–496.
26. *Губанов Д.А., Новиков Д.А., Чхартишвили А.Г.* Социальные сети: модели информационного влияния, управления и противоборства // М.: Издательство физико-математической литературы: МЦНМО. 2010. 228 с.
27. *Watts D., Dodds P.* Influentials, Networks, and Public Opinion Formation // Journal of Consumer Reseach. 2007. vol. 34. pp. 441–458.

References

1. Mitin N.A. *Novye modeli matematicheskoj psihologii i informacionnye process* [New models of mathematical psychology and information processes]. M.: IPM RAN. 2013. (In Russ.).
2. Social'naja set' [Social network]. Available at: https://ru.wikipedia.org/ (accessed: 12.01.2015). (In Russ.).
3. *Otchet o NIR «Issledovanie i obosnovanie principov korporativnoj jetiki v social'nyh setjah dlja voennosluzhashhih i grazhdanskogo personala VS RF»* [The research report "Research and justification of the principles of corporate ethics in social networks for military and civilian personnel of the armed forces". Part. 1]. SPb.: VKA imeni A.F. Mozhajskogo. 2014. 165 p. (In Russ.).
4. Korshunov A.V. [Objectives and methods of determining the attributes of the users of social networks] *Trudy 15-j Vserossijskoj nauchnoj konferencii «Jelektronnye biblioteki: perspektivnye metody i tehnologii, jelektronnye kollekcii»* [Proceedings of the 15th All-Russian Scientific Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" (RCDL'2013)]. 2013. pp. 183–193. Available at: http://ceur-ws.org/Vol-1108/paper23.pdf (accessed: 13.10.2014). (In Russ.).
5. Ogneva E.A. *Kognitivnoe modelirovanie konceptosfery hudozhestvennogo teksta* [Cognitive modeling concept sphere of artistic text. 2-nd edition]. Moscow: Editus, 2013. 282 p. (In Russ.).
6. Semashko K.V., Sheremet I.A. *Matematicheskoe modelirovanie informacionno-psihologicheskih otnoshenij v sociumah* [Mathematical modeling of information-psychological relations in societies]. Moscow: Nauka, 2007. 157 p. (In Russ.).
7. ABBYY Intelligent Search SDK. Available at: http://www.abbyy.ru/isearch/compreno/ (accessed: 23.12.2014).
8. ABBYY Compreno. Available at: http://www.dialog-21.ru/digests/dialog2012/materials/ pdf/ anisimovich.pdf (accessed: 23.12.2014).
9. ABBYY Intelligent Tagger SDK. Available at: http://www.abbyy.ru/adx/aspx/ adxgetmedia.aspx ?DocID=f6bad99e-d66a-4da0-9112-4c6fc15e1f72 (accessed: 24.12.2014).
10. Anisimovich K.V., Druskin K.Y., Zuev, K.A., Minlos F.R., Petrova M.A., Selegey V.P. [Syntactic and semantic parser based on linguistic technologies].

Mezhdunarodnaja konferencija po komp'juternoj lingvistike «Dialog» [International Conference on Computational Linguistics "Dialogue"]. Available at: http://www.dialog-21.ru/digests/dialog2012/materials/pdf/Anisimovich.pdf (accessed: 24.12.2014). (In Russ.).

11. Ermakova L.M. [Methods of extracting information from the text]. *Vestnik Permskogo universiteta – Perm University Bulletin*. 2012. vol. 1(9). pp. 77–84. Available at: http://dspace.nsu.ru:8080/jspui/bitstream/nsu/202/1/278_84017.pdf (accessed: 29.12.2014). (In Russ.).
12. Leonov Yu C., Fedotov A. M. [The Extraction of knowledge and facts from the texts of theses and abstracts for studying the relationship between scientific communities]. *Trudy 15-j Vserossijskoj nauchnoj konferencii «Jelektronnye bib-lioteki: perspektivnye metody i tehnologii, jelektronnye kollekcii»* [Proceedings of the 15th All-Russian Scientific Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" (RCDL'2013)]. 2013. pp. 32–41. Available at: http://ceur-ws.org/Vol-1108/paper5.pdf (accessed: 13.10.2014). (In Russ.).
13. Lukashevich N.V. [The extraction of knowledge and facts from texts]. *Nauchno-issledovatel'skij vychislitel'nyj centr MGU im. M.V.Lomonosova – Research computing center of M.V. Lomonosov Moscow State University*. Available at: http://www.slideshare.net/msucsai/2007-12 (accessed: 13.10.2014). (In Russ.).
14. Atyagina A.P. [Twitter as a new discursive practice on the Internet]. *Vestnik Omskogo universiteta - Bulletin of Omsk University*. 2012. vol. 4. pp. 203–208. (In Russ.).
15. Korshunov A.V. [Extract key terms from the microblogging messages via Wikipedia]. *Trudy Instituta sistemnogo programmirovanija RAN – Proceedings of Institute for System Programming RAS*. 2011. vol. 20. pp. 269–282. (In Russ.).
16. Mol' A. *Sociodinamika kul'tury* [Sociodynamics culture]. Moscow: Publ. «Progress». 1973. 405 p. (In Russ.).
17. Golovin S.Y. *Slovar' prakticheskogo psihologa* [Dictionary of practical psychologist]. Minsk: Harvest. 1998. 800 p. (In Russ.).
18. Festinger L. Theory of Cognitive Dissonance. Standford. CA. Standford University Press. 1957.
19. Heider F. The psychology of interpersonal relations. N.Y. 1958.
20. Newcomb T.M. An approach to the study of communicative acts. *Psychological Review*. 1953. vol. 60. pp. 393–404.
21. Osgood C.E., Tannenbaum P. The priciple of congruity in the prediction of attitude change. *Psychological review*. 1955. vol. 62. pp. 42-55.
22. Osgood C.E., Suci G. Tannenbaum P. The measurement of meaning. Chicago. «Semantic differential technique». 1957. 342 p.
23. Bartlett F. [Man remembers: Readings in the psychology of memory*]. Hrestomatija po psihologii pamjati – Readings on the psychology of memory*. Moscow: Publ. MGU. 2002. pp. 292–303. (In Russ.).
24. Ermakov Y.A. *Manipuljacija lichnost'ju: smysl, priemy, posledstvija* [Manipulation personality: meaning, methods, consequences]. Ekaterinburg: Izdatel'stvo Ural'skogo universiteta, 1995. 136 p. (In Russ.).
25. Pilkevich S.V., Lomako A.G. [Models of communication, while protecting from negative information influence]. *Metody obespechenija informacionnoj kiberbezopasnosti - Methods of providing information cybersecurity*: Moscow: KomKniga, 2013. vol 27. pp. 475–496. (In Russ.).
26. Gubanov D.A., Novikov D.A., Chkhartishvili A.G., *Social'nye seti: modeli informacionnogo vlijanija, upravlenija i protivoborstva* [Social networks: model information influence, control and confrontation]. Moscow: Izdatel'stvo fiziko-matematicheskoj literatury: MCNMO. 2010. 228 p. (In Russ.).

27. Watts D., Dodds P. Influentials, Networks, and Public Opinion Formation. *Journal of Consumer Research*. 2007. vol. 34. pp. 441–458.

Пилькевич Сергей Владимирович — к-т техн. наук, докторант кафедры систем сбора и обработки информации, Военно-космическая академия имени А.Ф. Можайского. Область научных интересов: информационная безопасность, криптография, моделирование социальных систем. Число научных публикаций — 60. ambers@list.ru; ул. Ждановская, д. 13, Санкт-Петербург, 197198; р.т.: +7(812) 237-19-60.

Pilkevich Sergey Vladimirovich — Ph.D., doctoral student of system for collecting and processing information department, Mozhaisky Military Space Academy. Research interests: information security, cryptography, modeling social systems. The number of publications — 60. ambers@list.ru; 13, Zhdanovskaya street, St.-Petersburg, 197198, Russia; office phone: +7(812) 237-19-60.

Еремеев Михаил Алексеевич — д-р техн. наук, начальник кафедры систем сбора и обработки информации, Военно-космическая академия имени А. Ф. Можайского. Область научных интересов: информационная безопасность, криптография, моделирование конфликтующих систем, автоматизированные системы сбора и обработки информации. Число научных публикаций — 200. mae1@rambler.ru; ул. Ждановская, д. 13,Санкт-Петербург, 197198; р.т.: +7(812) 237-19-60.

Eremeev Mikhail Alekseevich — Ph.D., Dr. Sci., head of system for collecting and processing information department, Mozhaisky Military Space Academy. Research interests: information security, cryptography, modeling of the conflicting systems. The number of publications — 200. mae1@rambler.ru; 13, Zhdanovskaya street, St.-Petersburg, 197198, Russia; office phone: +7(812) 237-19-60.

РЕФЕРАТ

Пилькевич С.В., Еремеев М.А. **Модель социально значимых Интернет-ресурсов.**

Стремительное развитие информационных технологий и появление новых средств массовой коммуникации многократно усилили возможности дистанционного взаимодействия отдельных пользователей информационно-телекоммуникационных систем, больших групп людей и населения страны в целом.

Появление социально значимых Интернет-ресурсов (социальные сети, форумы, онлайн-дневники, микро-блоги) стимулировало развитие новых способов выражения общественного мнения, которое оказывает влияние на государственные решения, задет основной вектор социального развития и поведения общества, влияет на интенсивность проявления конформизма, изменяет мнение социума по важнейшим, экономическим и духовным вопросам.

Публикация посвящена вопросу разработки модели социально значимых Интернет-ресурсов. Рассматриваемая модель состоит из трех взаимосвязанных моделей: модели пользовательского аккаунта, модели информации, генерируемой пользователями социально значимых Интернет-ресурсов, а также модели информационно-психологического отношения *B*. Описание элементов предлагаемой модели представлено с помощью теоретико-множественного аппарата и теории графов, для формализации функциональных зависимостей использованы подходы апробированные в рамках теории коммуникаций, теории информации, теории когнитивного соответствия, физиологии и социальной психологии.

Применение междисциплинарного подхода к моделированию социально значимых Интернет-ресурсов позволило отразить основные системные свойства современного «информатизированного» социума.

Разработанная модель социально значимых Интернет-ресурсов позволяет в равной степени адекватно описывать информационно-психологическое взаимодействие участников основных современных видов социальных ресурсов: форумов, социальных сетей общения, блогов и Интернет СМИ.

SUMMARY

Pilkevich S.V., Eremeev M.A. **Model of Socially Important Internet Resources.**

The rapid development of information technologies and introduction of new means of mass communication increased the possibility of remote interaction between the user of information and telecommunication systems, large groups of people and the population as a whole.

The emergence of socially significant Internet resources (social networks, forums, online blogs, micro-blogs) stimulated the development of new ways of expressing public opinion, which has an impact on government decisions that affected the main vector of social development and behavior society, affects the intensity of conformism, changes the opinion of the society on the most important economic and spiritual matters.

The publication is dedicated to developing a model of socially important Internet resources. This model consists of three interrelated models: models of the user account, the model of information generated by users of socially significant by Internet resources, as well as models of information-psychological relations *B.* description of the elements of the proposed model is presented using a set-theoretic apparatus and graph theory, formalization of functional dependencies used approaches tested in the framework of the theory of communications, information theory, cognitive theory of conformity, the physiological and social psychology.

The use of an interdisciplinary approach to the modeling of socially important Internet resources allowed to reflect the core system properties modern "computerized" society.

The developed model of socially important Internet resources allows equally adequately describe the information-psychological interaction between the participants and the main modern types of social resources: forums, social networks, communication, blogs and online media.

УДК 681.3.067

И.А. Носаль

МЕТОД ОБОСНОВАНИЯ МЕРОПРИЯТИЙ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ СОЦИАЛЬНО-ВАЖНЫХ ОБЪЕКТОВ

Носаль И.А. **Метод обоснования мероприятий информационной безопасности социально-важных объектов.**

Аннотация. Рассматривается задача обоснования целесообразных мероприятий информационной безопасности социально важных объектов. Для ее решения предлагается усовершенствованный метод, ориентированный на более гибкий учет особенностей текущих ситуаций. Предложен ряд новых марковских моделей защищаемых процессов и возможных угроз применительно к структурам Пенсионного фонда. Приведены результаты моделирования.

Ключевые слова: метод, модели, информационная безопасность, обоснование мероприятий.

Nosal I.A. **Method of Information Security Measures Substantiation for Socially Important Objects.**

Abstract. The justification of appropriate information security measures of socially important objects is considered. To solve it an improved method that focuses on a more flexible accounting features of the current situation is provided. The set of new Markov models protected processes and possible threats in relation to the structures of the Pension Fund is offered. The simulation results are given.

Keywords: method, models, information security, substantiation and feasibility of measures.

1. Введение. Под социально-важным объектом (СВО) в данной работе подразумевается социально-ответственный институт, не являющийся при этом органом государственной власти. Основной целью СВО является предоставление государственных услуг населению (обеспечение прав граждан), нарушение или прерывание работы которого может привести к нарушению нормальной жизнедеятельности населения.

Примеры таких организаций – Фонд социального страхования, Фонд обязательного медицинского страхования, Пенсионный фонд Российской Федерации. Все эти организации по форме образования и расходования денежных средств являются внебюджетными государственными фондами, имеют схожие цели, задачи, принципы работы, административно-управленческую структуру и являются крупнейшими операторами персональных данных.

Информационная безопасность для социально-важных объектов – это одно из главнейших условий надлежащего предоставления ими качественных государственных услуг населению.

Она является частью системы национальной безопасности и внутренней политики, а также влияет на безопасность личности, общества и государства [1]. Для государства обеспечение информационной безопасности СВО гарантия надлежащего исполнения своих функций (обязательств перед населением).

Для обеспечения информационной безопасности (ИБ) СВО предоставляется не много инструментов и ресурсов, однако предъявляется достаточно требований, как со стороны государства, так и со стороны населения. И, поскольку, каждый субъект этой системы имеет свой круг интересов и задач, которые он решает с помощью СВО, а значит разнятся наборы требований, которые предъявляются к ИБ СВО. Поэтому система информационной безопасности СВО остро нуждается в удобном, эффективном и надежном методе поиска и обоснования мероприятий ИБ. Необходимо также разработать систему моделей угроз информационной безопасности актуальных для деловых процессов СВО, учитывая широкий круг условий, отражающих объективные закономерности.

Известны многочисленные экономические и математические методы решения этой задачи [2–8]. В большинстве случаев применяются методы, базирующиеся на статистических данных и экспертных оценках с применением различного математического аппарата [9–11]. Но даже методы, использующие хорошо адаптированные для этих целей математические инструменты оценки, такие как нечеткие множества, нечеткая логика и искусственные нейронные сети [12–15], в конечном счете, опираются на опыт и субъективные мнения экспертов. Это влечет за собой все недостатки и проблемы использования экспертных методов оценивания [16–18].

Следует также упомянуть о требованиях государственных регуляторов в области ИБ [19–23], отраслевых и международных стандартах ИБ [24–27]. Все они в той или иной мере раскрывают подходы, которыми должна руководствоваться организация при выборе мероприятий ИБ. Чаще предлагают, основываясь на предложенной методике, выбрать из ограниченного списка соответствующие контрмеры, а иные и вовсе тоталитарно требуют выполнения конкретных мер. В итоге, при отсутствии высококвалифицированных специалистов, которые смогут эффективно и с пониманием внедрить требования стандарта в существующую систему, обоснование мероприятий ИБ сводится к тому, что их выполнение обязательно, со всеми вытекающими последствиями. К эти последствиям относятся: нарушения или затруднения деятельности СВО, увеличение нагрузки на персонал, усложнение документооборота, дублирование документации, мер и методов защиты. Как показали результаты расчетов, опубликованные в работе [28], даже «подогнанный» ISO/IEC 27001 – 2006 (универсальный стандарт по ИБ для организаций любых типов, размеров, отраслей) не отражает всей картины, не охватывает

уникальные для конкретной организации детали и не гарантирует адекватность и обоснованность выстроенной системы защиты.

Главный недостаток всех перечисленных методов в том, что область их использования ограничена и распространяется либо на противодействие именно техническим и сетевым атакам, либо основывается исключительно на оценке экономической эффективности мероприятий. В первом случае объектом защиты является информационная система. Не рассматривается безопасность всего объекта информатизации, непрерывность деловых процессов, не учитываются интересы владельцев бизнеса и информации. Во втором случае не учитывается широкий ряд других важных параметров: удобство для пользователей, интегрируемость в текущую инфраструктуру, контролируемость и т.п. Выстраиваемая таким образом система обеспечения ИБ не в полной мере обладает требуемыми свойствами: системностью, интегрируемостью, комплексностью, прозрачностью, адекватностью, оптимальностью и подконтрольностью.

Поэтому необходимо совершенствование соответствующего научно-методического аппарата. Требуется разработка более точного метода обоснования целесообразности принятия решений и мероприятий ИБ, учитывающего особенности объекта защиты, его деловых процессов, внешней и внутренней среды функционирования, позволяющего повысить уровень информационной безопасности СВО в целом.

Необходимо решить научно-техническую задачу по разработке новых моделей и методов обоснования мероприятий информационной безопасности СВО, повышающих эффективность функционирования этих объектов.

2. Метод гибкого обоснования мероприятий ИБ. Учитывая ранее полученные результаты [29, 30] в интересах решения сформулированной задачи предлагается уточненный метод гибкого обоснования мероприятий ИБ СВО.

В обобщенном виде этот метод можно представить в виде следующей последовательности шагов:

Ш1. Анализ защищаемого процесса, выделение ключевых его особенностей.

Ш2. Анализ текущего состояния ИБ на защищаемых участках, уточнение или пересмотр целей защиты информации, условий их достижения.

Ш3. Разработка нескольких альтернативных моделей защищаемого процесса с учетом того или иного набора мероприятий

ИБ. В частности, опираясь на предельную теорему теории вероятностей для потоков событий, защищаемый процесс можно рассматривать как марковский. В этом случае такие модели могут быть представлены в виде соответствующих процессу графов состояний.

Ш4. Формулировка условий оценивания эффективности мероприятий ИБ (требований к итоговой безопасности процесса, входным и выходным экономическим показателям, ограничений по времени выполнения мероприятия или времени простоя и т.п.) на качественном уровне.

Ш5. Разработка адекватной оптимизационной модели ИБ.

Ш6. Определение текущих параметров переходов моделируемого процесса из одних состояний в другие.

Ш7. Задание начальных и интересующих состояний процесса.

Ш8. Расчет вероятностей нахождения процесса в интересующих состояниях для каждого из альтернативных наборов мероприятий и определение значений других показателей эффективности, входящих в выбранную оптимизационную модель. В качестве таких показателей могут выступать временные и материальные затраты, интегральные потери ценности защищаемой информации на заданном интервале времени, математическое ожидание удовлетворенных заявок и другие.

Ш9. Проверка выполнимости условий, связанных с этими показателями.

Ш10. Поиск экстремума основного показателя эффективности (целевой функции) на заданном наборе альтернативных мероприятий, удовлетворяющих условиям задачи.

Ш11. Принятие в качестве целесообразного того мероприятия, при котором достигнут экстремум целевой функции.

При разработке марковской модели защищаемого процесса с учетом мероприятий ИБ в виде графа состояний следует исходить из целесообразного уровня формализации этого процесса. Излишняя детализация влечет за собой повышение затрат на разработку модели процесса и определение ее параметров. Грубая формализация позволяет оперативно получать интересующие оценки, однако не обеспечивает необходимой точности результатов. Для определения целесообразного уровня формализации анализируемых процессов применим метод экспертных оценок.

Для построения такого графа, в соответствии с предлагаемым подходом, следует, прежде всего, определиться с уровнем масштабирования модели, интересующими результатами и изучить объект моделирования – защищаемый деловой процесс, затем

выделить:

– основные этапы его выполнения;

– задействованные ресурсы;

– определить есть ли в процессе стандартные ответвления, связанные с принятием решений;

– какие могут быть ошибки, сбои в выполнении процесса;

– какие атаки на процесс актуальны и к каким последствиям (нарушениям) на каких этапах могут привести;

– какие из нарушений связаны с защищаемыми ресурсами, какие с нарушениями требований, какие с человеческими ошибками.

На основании указанных выше данных могут быть выделены последовательности состояний, в которых может находиться процесс, и возможные переходы между ними. Следует уделить построению модели процесса в виде графа состояний наибольшее внимание, поскольку именно этим выбором будет определяться область получаемых результатов и характер возможных выходных данных.

Следующим важным моментом является задание исходных данных, условий поиска. Одновременно в качестве них могут выступать следующие параметры:

– наличие тех или иных связей и переходов из состояния в состояние;

– интенсивности переходов из состояния в состояние;

– вероятности нахождения в рассматриваемых состояниях на момент $t = 0$;

– затраты на реализацию защиты;

– ущерб от реализации угрозы и другие.

В качестве входных данных интенсивности переходов из состояния в состояние и ограничения по времени могут задаваться, основываясь на регламенте моделируемого делового процесса, статистических оценках и требуемых значениях показателей.

Некоторые из интенсивностей переходов зависят от особенностей мероприятий ИБ (к примеру, от частоты проведения проверок). Эта неопределенность может быть исключена в ходе разработки оптимизационных моделей ИБ, поскольку эти интенсивности становятся искомыми параметрами. В другом случае, они могут быть определены путем сбора и обработки статистических данных. В ряде случаев, когда известны начальные и конечные состояния процесса на некотором интервале времени определение исходных интенсивностей осуществимо также путем подбора параметров с использованием метода наименьших квадратов.

Важным моментом моделирования является распознавание

состояний, в которых система может находиться в исходный момент времени.

Для расчета вероятностей нахождения процесса в интересующих состояниях в соответствии с построенным графом составляется система дифференциальных уравнений. Затем она разрешается относительно заданных начальных и интересующих состояний.

В соответствии с этим методом выбор конкретной оптимизационной модели должен осуществляться, исходя из наибольшего соответствия ее реальной ситуации с учетом преследуемых целей, текущих условий и неопределенностей.

Новизна этого метода обоснования мероприятий ИБ СВО состоит в особенностях его отдельных этапов, в возможностях гибкого реагирования на возникающие ситуаций. При обосновании мероприятий ИБ СВО предлагается учитывать более широкий круг возможных условий - ограничений, свойственных не только ИБ СВО, но и самому процессу обоснования. При практическом использовании этого метода возможно накопление множества готовых для использования моделей целесообразных мероприятий ИБ, ориентированных на типовые ситуации.

3. Модели защищаемых процессов социально-важных объектов. Для реализации рассмотренного метода предлагается использовать следующие типовые модели защищаемых процессов, характерных для СВО, применительно к структурам Пенсионного фонда (рисунок 1). На рисунке 1*a* приняты обозначения: 1 – начальное состояние (поступление заявления); 2 – прием и проверка правильности оформления представленных документов; 3 – возврат заявления без регистрации и разъяснение причины отказа; 4 – регистрация заявления; 5 – заявление ошибочно зарегистрировано. Согласно рисунку 1*b* выделены: 1 – начальное состояние (направление запроса в другие органы (истребование документов); 2 – поступление необходимых документов; 3 – принято корректное решение об удовлетворении заявления/отказе в приеме заявления; 4 – документы не получены/получены поддельные документы, 5 - решение принято не верно). Графу на рисунке 1c свойственны состояния: 1 – принято решение об удовлетворении заявления; 2 – произведен расчет сумм, полагающихся к выплате в соответствии с последним принятым решением; 3 – выплатные документы сформированы; 4 – выплатные документы приняты на выплату; 5 – в выплатные документы внесены несанкционированные изменения, 6 – на выплату приняты выплатные документы с несанкционированными изменениями. На рисунке 1*d*

приняты обозначения: 1 – выплатные документы приняты на выплату; 2 – произведена проверка принятых на выплату документов; 3 – выплата произведена в соответствии с законодательством; 4 – выплата произведена некорректно (поддельные списки направлены в банк, платежные документы с не корректными суммами направлены в Казначейство).

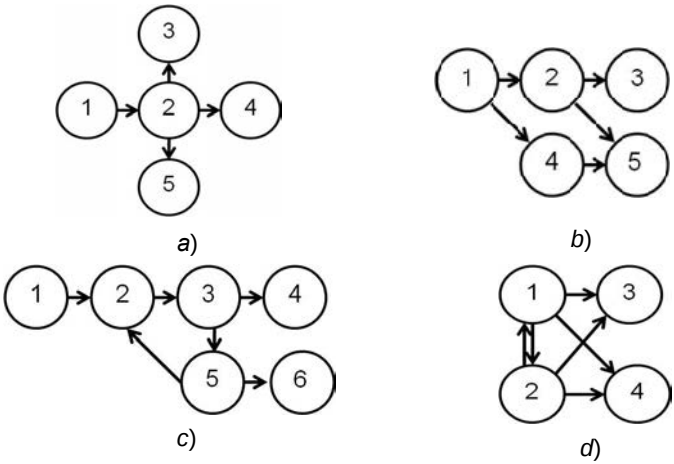


Рис. 1. Типовые модели защищаемых процессов в структурах Пенсионного фонда: *a* - модель приема документов; *b* - модель принятия решений по назначению пенсий; *c* - модель перерасчета выплат; *d* - модель осуществления выплат

Модели на рисунке 1 представлены в виде графов состояний. Дугам этих графов могут быть поставлены в соответствие интенсивности переходов из состояния в состояние. Они в свою очередь могут задаваться, исходя из регламента моделируемого делового процесса. Для каждой из этих моделей свойственна соответствующая система дифференциальных уравнений.

Поясним модели на рисунке 1. Согласно модели на рисунке 1*c*, чем меньше процесс находится в состоянии 6, тем выше уровень ИБ. Также уровень ИБ зависит от изменения параметров переходов из состояния 3 в состояния 4 и 5, и из состояния 5 в состояния 2 и 6. Увеличение времени перехода из состояния 5 в состояние 6 связано с улучшением эффективности защитных мер. Заметим, что в случае предоставления пользователям неоптимального набора прав доступа, граф состояний может обладать еще одним ребром перехода из состояния 2 в состояние 5, поскольку несанкционированные

изменения могут оказаться не исправленными.

Точно такая же ситуация с моделью на рисунке 1*c*. Актуальной угрозой для этого процесса является возможность некорректной проверки направленных на выплату документов и отправки документов с поддельными данными. Вероятность этой угрозы можно снизить только контролем выдачи ролей и прав доступа на ввод и проверку данных. В случае, когда процесс организован по всем требованиям безопасности, то отсутствуют какие-либо искаженные данные. Переход из состояния 2 в состояние 4 исключен и наоборот. При этом, уровень ИБ зависит от изменения параметров переходов в состояния 3 и 4.

Новизна этих моделей состоит в структурах предложенных графов состояний, отражающих объективные закономерности формализуемых процессов, применительно к Пенсионному фонду.

Представленные на рисунке 1 модели могут быть использованы в качестве типовых при обосновании целесообразных мероприятий ИБ на СВО. Однако кроме них необходимо иметь также модели процессов нарушения ИБ СВО.

4. Модели процессов нарушения ИБ СВО. В качестве таких типовых моделей могут выступать графы состояний, показанные на рисунке 2.

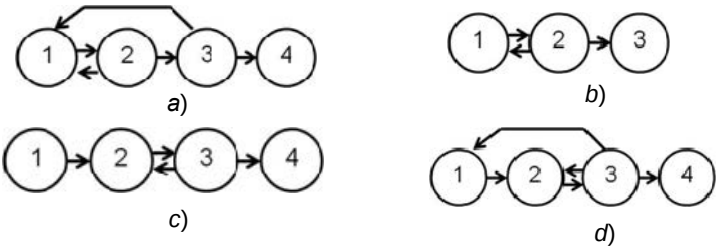


Рис. 2. Типовые модели нарушения информационной безопасности, применительно к структурам Пенсионного фонда: *a* - модель подмены источника (поставщика) данных, атака «masquerading»; *b* - модель перехвата передаваемых данных, атака «man in the middle» (компрометация канала связи); *c* – модель нарушения целостности данных; *d* - модель подачи поддельных данных на входе

В качестве вершин графа на рисунке 2*a* выделены: 1 – злоумышленник отследил обращения СВО к внешнему источнику (поставщику) данных; 2 – запрос СВО к внешнему источнику данных перехвачен; 3 – злоумышленник замаскировался и направил ответ от имени источника; 4 – некорректные данные

приняты (нарушение аутентичности и целостности данных). На рисунке 2*b*: 1 – злоумышленник отследил обращения СВО к внешнему источнику (поставщику) данных; 2 – злоумышленник перехватил/вычислил/подобрал ключевую/парольную информацию; 3 – НСД к защищаемым информационным ресурсам получен. Согласно рисунку 2*c*: 1 – злоумышленник получил доступ к данным; 2 – злоумышленник внес изменения в данные; 3 – данные проверены на подлинность; 4 – некорректные данные приняты. Для модели на рисунке 2*d* характерны следующие состояния процесса: 1 – злоумышленником сформированы поддельные документы; 2 – поддельные документы поданы в СВО; 3 –поддельные документы приняты СВО; 4 – деловой процесс нарушен.

Согласно рисунку 2 применительно к структурам Пенсионного фонда, как СВО, неправомерное начисление пенсии из-за нарушений ИБ может произойти:

- на уровне персонифицированного учета за счет подачи в СВО поддельных документов, на протяжении всей жизни;

- путем подделки дополнительных документов, влияющих на социальные надбавки независимые от пенсионного капитала;

- внесением изменений непосредственно в суммы выплаты пенсии и другими способами.

Возможны четыре типа атак на процесс назначения и выплаты пенсий (НВП). Первая – подделка документов еще до момента их подачи в СВО, таким образом, чтобы создать эффект «мертвых душ». Тогда ни на одном из этапов невозможно будет обнаружить, что такого человека фактически не существует. Данную угрозу реализует внешний нарушитель. Здесь идет нарушение аутентичности информации (подмена персональных данных – одного человека выдают за другого);

Вторая – подделка документов на этапе проверки правильности заявления и сопутствующих документов, когда оператор принимает к регистрации оформленные с нарушениями документы. Это может быть как заявитель, так и ошибка или содействие оператора, принимающего документы.

Третья – подделка документов на этапе ответа на запрос, когда в СВО приходят искаженные данные о лице, подавшем заявление, что провоцирует неверное определение прав на предоставление госуслуги и неправомерное принятие решения об удовлетворении или отказе в услуге. Эта угроза может осуществляться как внутренними нарушителями (с которыми реализуется взаимодействие в рамках оказания услуги, инсайдеров в СВО), так внешним

злоумышленниками с помощью технического перехвата данных – атака “man in the middle”.

Четвертая – это неправомерная корректировка сумм выплат перед направлением списков в банки. Последняя в этом списке, но первая по количеству инцидентов атака, реализуется внутренним нарушителем. К данной категории нарушений следует отнести также подделку данных (не только сумм выплат, а любых данных) уже непосредственно при передаче в банки и расчетный центр с использованием атаки "man in the middle”.

Общими, для процессов обеспечения ИБ являются противоречия между: уровнем защищенности и доступности информационных ресурсов; уровнем защищенности и затратами на обеспечение ИБ; затратами на обеспечение ИБ и возможным информационным ущербом со стороны несанкционированных пользователей и другие.

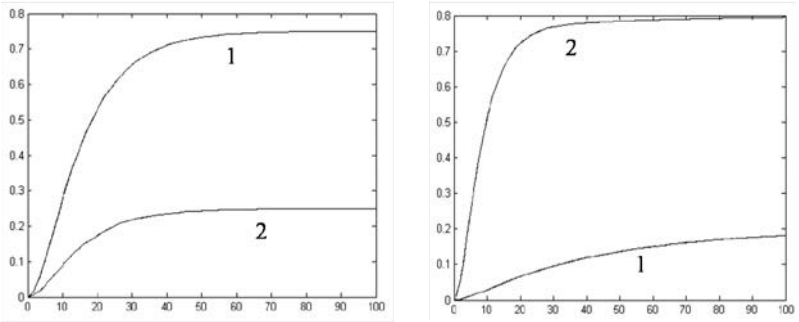
Предложенные модели нарушения ИБ СВО расширяют взгляды на возможные угрозы процессам, свойственным структурам Пенсионного фонда.

5. Результаты моделирования. В интересах подтверждения справедливости предложенных теоретических положений по обоснованию ИБ на СВО проводилось математическое моделирование. Исходные данные по параметрам моделей, отраженных на рисунках 1, 2, формировались автором с учетом ранее накопленного опыта практической деятельности в сфере ИБ на СВО.

В результате такого моделирования с применением пакета прикладных программ MatLab применительно к графам на рисунках 1*b* и 1*c* , были получены зависимости, приведенные на рисунках 3*a*, *b* и 4*a*, *b*, соответственно.

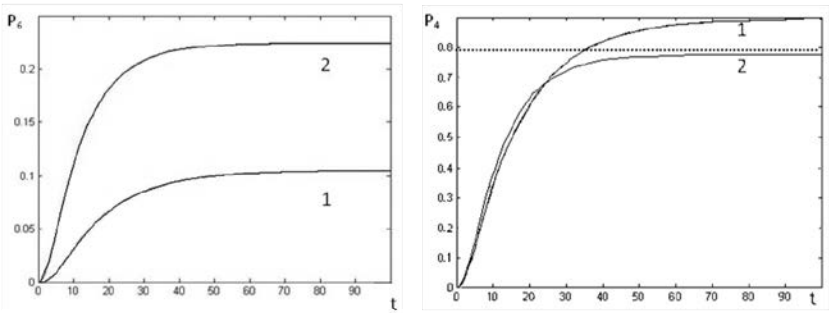
Для анализа был выбран конечный промежуток времени от нуля до 100 минут, поскольку в дальнейшем функции не меняли своего поведения. Все рассчеты были произведены при начальных состояниях [1,0,0,0,0].

Сравнивая рисунки 3*a* *и* 3*b*, наблюдаем прямо пропорциональную картину изменения вероятностей принятия правильного и не правильного решений в зависимости от наличия и отсутствия защищаемых информационных ресурсов (ЗИР). Эти графики наглядно показывают насколько важна роль ЗИР, достоверных входных данных, документов, необходимых для процесса назначения и расчета пенсии.



a) *b*)

Рис. 3. Зависимость вероятности корректного (кривая 1) и некорректного (кривая 2) принятия решений от времени при наличии (*a*) и отсутствии защищаемых информационных ресурсов (*b*)



a) *b*)

Рис. 4. Зависимости вероятности некорректной (a) и корректной (b) выплаты пенсии от времени для наборов мероприятий защиты с разделением прав (кривая 1) доступа и без их разделения (кривая 2)

Согласно рисункам 3 а,б на момент времени, равный 20-ти минутам, вероятность принятия верного решения равна 0,5282 и 0,06561, соответственно. Таким образом, можно говорить о том, что правильно полученные документы, которые находятся в ведомстве других организаций, имеют вклад, оказываемый на правильность принятого решения, равный 46%. Интересно также наблюдать характер роста вероятности принятия правильного решения в обоих вариантах. В первом случае первые 30 минут наблюдается экспоненциальный рост. Последующие 0,5 часа рост продолжается, но значительно медленнее, и практически прекращается на 60-той минуте. В итоге достигается значение вероятности, равное 0,8. Во втором случае вероятность принятия верного решения возрастает

постоянно с практически неизменной скоростью и на наблюдаемом промежутке не превышает 0,2.

Для примера обоснования мероприятий ИБ применительно к процессам Пенсионного фонда в соответствии с моделью в виде графа состояний на рисунке 1*с* характерны несколько другие зависимости от времени вероятностей нахождения его в состояниях P_6 и P_4.

Из анализа рисунка 4 видно, что мероприятия защиты с разделением прав доступа являются более эффективными, чем со смешением этих прав.

Оптимизацию мероприятий ИБ, в частном случае, можно осуществлять, исходя из минимизации вероятности некорректной выплаты пенсий при различных дополнительных условиях – ограничениях.

Для расчета ценности защищаемых информационных ресурсов в рамках предлагаемого метода можно использовать известную модель [30].

Таким образом, результаты моделирования показывают, что предлагаемые решения позволяют успешно обосновывать мероприятия ИБ.

7. Заключение. Для обоснования мероприятий ИБ СВО предлагается более полно учитывать особенности текущих ситуаций. В соответствии с этими ситуациями рекомендуется гибко разрабатывать модели целесообразных мероприятий ИБ. Отличительной особенностью предлагаемого метода выступает ориентированность его на использование для формализации ИБ СВО математического аппарата марковских процессов. Учитывая, что на все случаи невозможно заранее разработать целесообразные модели мероприятий ИБ СВО, следует опираться на комплексы типовых моделей, которые при необходимости можно легко корректировать. При разработке марковских моделей исследуемых процессов следует исходить из целесообразного уровня их формализации. Излишняя детализация влечет за собой повышение затрат на разработку модели и определение ее параметров. Грубая формализация позволяет оперативно получать интересующие оценки, однако не обеспечивает необходимой точности результатов.

Предложенные в статье новые типовые марковские модели защищаемых процессов и возможных угроз могут быть успешно применимы в практической деятельности специалистов по защите информации при обосновании целесообразных мероприятий ИБ СВО.

Литература

1. *Юсупов Р.М.* Наука и национальная безопасность. 2-е издание, переработанное и

дополненное // СПб.: Наука. 2011. 369 с.
2. *Петренко С.А., Попов Ю.И.* Оценка затрат на информационную безопасность // Конфидент. 2003. №1(49). С. 68–73.
3. *Обухов А.А.* Диагностика необходимости инвестирования в безопасность в современном предпринимательстве и формирование на ее основе рекомендаций // Вестник Омского университета. Серия «Экономика». 2013. №3. С. 78–84.
4. *Рытов М.Ю., Рудановский М.В.* Управление безопасностью информационных технологий на основе методов когнитивного моделирования // Информационная безопасность. 2010. №4. С. 579–582.
5. *Ажмухамедов И.М., Ханжина Т.Б.* Оценка экономической эффективности мер по обеспечению информационной безопасности // Вестник АГТУ. 2011. № 1 С. 185–190.
6. *Бирюков Д.Н., Ломако А.Г.* Подход к построению системы предотвращения киберугроз // Проблемы информационной безопасности. Компьютерные системы. 2013. №2. С. 13–19.
7. *Крамаров Л.С., Бабенко Л.К.* Обнаружение сетевых атак и выбор контрмер в облачных системах // Известия ЮФУ. Технические науки. 2013. №12(149). С. 94–101.
8. *Абрамов Е.С., Кобилев М.А., Крамаров Л.С., Мордвин Д.В.* Использование графа атак для автоматизированного расчета мер противодействия угрозам информационной безопасности сети // Известия ЮФУ. Технические науки. 2014. №2(151). С. 92–100.
9. *Трофимов И.Б.* Методы оценки информационной безопасности предприятия на основе процессного подхода:дис. канд. техн. наук // Санкт-Петербург. 2010. 134 с.
10. *Ажмухамедов И.М., Ханжина Т.Б.* Оценка экономической эффективности мер по обеспечению информационной безопасности // Вестник АГТУ. 2011. №1. С. 185–190.
11. *Миронов В. В., Носаль И.А.* Моделирование и оценка системы обеспечения информационной безопасности на примере ГОУ ВПО «СыктГУ» // Информация и безопасность. 2011. № 2. С. 209–211.
12. *Васильев В.И., Савина И.А., Шарипова И.И.* Построение нечетких когнитивных карт для анализа и управления информационными рисками вуза // Вестник УГАТУ. 2008. №2. Т. 10. С. 199–209.
13. *Schneier B.* Attack Trees // Dr. Dobb's Journal. 1999. vol. 24. no. 12. pp. 21–29.
14. *Sodiya A. S., Onashoga S. A., Oladunjoye B. A.* Threat Modeling Using Fuzzy Logic Paradigm // Issues in Informing Science and Information Technology. 2007. vol. 4. pp. 53–61.
15. *Чечулин А.А.* Методика оперативного построения, модификации и анализа деревьев атак // Труды СПИИРАН. СПб: Наука. 2013. №3 (26). С.40–55.
16. *Карпеев Д.О.* Исследование и развитие методического обеспечения оценки и управления рисками информационных систем на основе интересо-ориентированного подхода: дис. канд. техн. наук // Воронеж. 2009. 171 с.
17. *Корнилова А.Ю,. Палей Т.Ф.* Проблемы применения методов экспертных оценок в процессе экономического прогнозирования развития предприятия // Проблемы современной экономики. 2010. № 3 (35). С.124–128.
18. *Ефимов Е.И.* Возможность применения существующих средств анализа рисков в системах принятия решений с привлечением экспертов // Омский научный вестник. 2011. № 3-103. С.281–284.
19. Методические рекомендации по обеспечению с помощью криптосредств безопасности персональных данных при их обработке в информационных системах персональных данных с использованием средств автоматизации. 2008. № 149/54-144.
20. Приказ Федеральной службы по техническому и экспортному контролю (ФСТЭК России) «Об утверждении Состава и содержания организационных и технических мер по обеспечению безопасности персональных данных при их

обработке в информационных системах персональных данных». 2013. № 21.
21. Приказ ФСТЭК России «Об утверждении состава и содержания организационных и технических мер по обеспечению безопасности персональных данных при их обработке в информационных системах персональных данных». 2013. № 7.
22. Положение Центрального Банка Российской Федерации «О требованиях к обеспечению защиты информации при осуществлении переводов денежных средств и о порядке осуществления Банком России контроля за соблюдением требований к обеспечению защиты информации при осуществлении переводов денежных средств». 2012. № 382-П.
23. Постановление Правительства Российской Федерации «Об утверждении Положения о защите информации в платежной системе». 2012. № 584.
24. Payment Card Industry Data Security Standard (PCI DSS) // PCI Security Standards Council LLC. Version 2.0. 2010. 75 p. URL: https://www.pcisecuritystandards.org/documents.
25. Стандарт Банка России СТО БР ИББС-1.0-2014 Обеспечение информационной безопасности организаций банковской системы Российской Федерации. Общие положения // М.:Вестник Банка России. 2014. № 48–49. 37 с.
26. ГОСТ Р ИСО/МЭК 27001—2006 Информационная технология. Методы и средства обеспечения безопасности. Системы менеджмента информационной безопасности. Требования // М.: Стандартинформ. 2008. 26 с.
27. ГОСТ Р ИСО/МЭК 13335-1 – 2006 Информационная технология. Методы и средства обеспечения безопасности. Часть 1. Концепция и модели менеджмента безопасности информационных и телекоммуникационных технологий // М.: Стандартинформ. 2006. 23 с.
28. *Осипов В. Ю., Носаль И. А*. Обоснование периода пересмотра мероприятий по защите информации // Информационно-управляющие системы. 2014. № 1. С. 63–69.
29. *Осипов В.Ю., Носаль И.А*. Обоснование мероприятий информационной безопасности // Информационно-управляющие системы. 2013. № 2(63). С. 48–53.
30. *Носаль И.А*. Обоснование оптимального набора прав доступа // Комплексная защита объектов информатизации и измерительные технологии. Сб. научн. тр. Всероссийской научно-практической конф. с междунар. участ. Санкт-Петербург:Издательство Политехнического университета. 2014. С. 41–45.

References

1. Jusupov R.M. *Nauka i nacional'naja bezopasnost'. 2-e izdanie, pererabotannoe i dopolnennoe*. [Science and national security. 2nd edition, revised and enlarged]. Spb.: Nauka. 2011. 369 p. (In Russ.).
2. Petrenko S.A., Popov Ju.I. [Assessment of the cost of information security]. *Konfident – Confident*. 2003. no.1(49). pp. 68–73. (In Russ.).
3. Obuhov A.A. [Diagnosis of the need to invest in security in today's business and formation on its basis recommendations]. *Vestnik Omskogo universiteta. Serija «Jekonomika» - Bulletin of Omsk University. Series "Economy"*. 2013. no. 3. pp. 78–84. (In Russ.).
4. Rytov M.Ju., Rudanovskij M.V. [Security management of information technologies based on cognitive modeling methods]. *Informacionnaja bezopasnost' – Information Security*. 2010. no. 4. pp. 579–582. (In Russ.).
5. Azhmuhamedov I.M., Hanzhina T.B. [Evaluation of cost-effectiveness of information security]. *Vestnik AGTU – Bulletin AGTU*. 2011. no. 1. pp. 185–190. (In Russ.).
6. Birjukov D.N., Lomako A.G. [Approach to building systems to prevent cyber threats]. *Problemy informacionnoj bezopasnosti. Komp'juternye sistemy – Problems of information security. Computer systems*. 2013. no. 2. pp. 13–19. (In Russ.).
7. Kramarov L.S., Babenko L.K. [Detection of network attacks and countermeasures in the range of cloud systems]. *Izvestija JuFU. Tehnicheskie nauki – Proceedings of the JuFU. Technical sciences*. 2013. no. 12(149). pp. 94–101. (In Russ.).

8. Abramov E.S., Kobilev M.A., Kramarov L.S., Mordvin D.V. [Using the attack graph for automated calculation of countermeasures network information security threats]. *Izvestija JuFU. Tehnicheskie nauki – Proceedings of the JuFU. Technical sciences.* 2014. no. 2(151). pp. 92–100. (In Russ.).
9. Tronikov I.B. *Metody ocenki informacionnoj bezopasnosti predprijatija na osnove processnogo podhoda:dis. kand. tehn. nauk.* [Methods for assessing information security based on the process approach: Thesis. cand. tehn. Sciences]. Spb. 2010. 134 p. (In Russ.).
10. Azhmuhamedov I.M., Hanzhina T.B. [Evaluation of cost-effectiveness of information security]. *Vestnik AGTU – Bulletin AGTU.* 2011. no. 1. pp. 185–190. (In Russ.).
11. Mironov V.V., Nosal' I.A. [Modeling and assessing information security system on the example of GOU VPO "SyktGU"]. *Informacija i bezopasnost' – Information and Security.* 2011. no. 2. pp. 209–211. (In Russ.).
12. Vasil'ev V.I., Savina I.A., Sharipova I.I. [Construction of fuzzy cognitive maps for analysis and information risk management university]. *Vestnik UGTU – Bulletin UGTU.* 2008. vol. 2. Issue 10. pp. 199–209. (In Russ.).
13. Schneier B. Attack Trees. *Dr. Dobb's Journal*. 1999. vol. 24. no. 12. pp. 21–29.
14. Sodiya A. S., Onashoga S. A., Oladunjoye B. A. Threat Modeling Using Fuzzy Logic Paradigm. Issues in Informing Science and Information Technology. 2007. vol. 4. pp. 53–61.
15. Chechulin A.A. [Methodology operational formation, modification and analysis of attack trees]. *Trudy SPIIRAN – SPIIRAS Proceedings. Spb.: Nauka*. 2013. no. 3(26). pp. 40–53. (In Russ.).
16. Karpeev D.O. *Issledovanie i razvitie metodicheskogo obespechenija ocenki i upravlenija riskami informacionnyh sistem na osnove intereso-orientirovannogo podhoda: dis. kand. tehn. nauk.* [Research and development of methodological support risk assessment and management of information systems based on interests-oriented approach: Thesis. cand. tehn. Sciences.]. Voronezh. 2009. 171 p. (In Russ.).
17. Kornilova A.Ju,. Palej T.F. [Problems of application of expert assessments in the process of economic forecasting enterprise development]. *Problemy sovremennoj jekonomiki – Problems of the modern economy.* 2010. no. 3 (35). pp.124–128. (In Russ.).
18. Efimov E.I. [The possibility of using existing tools of risk analysis in decision-making systems with experts]. *Omskij nauchnyj vestnik – Omsk Scientific Bulletin.* 2011. no. 3. pp. 281–284. (In Russ.).
19. *Metodicheskie rekomendacii po obespecheniju s pomoshh'ju kriptosredstv bezopasnosti personal'nyh dannyh pri ih obrabotke v informacionnyh sistemah personal'nyh dannyh s ispol'zovaniem sredstv avtomatizacii* [Methodical recommendations for using kriptosredstv personal data security at their processing within the information systems of personal data with the use of automation]. 2008. No 149/54-144. (In Russ.).
20. *Prikaz Federal'noj sluzhby po tehnicheskomu i jeksportnomu kontrolju (FSTJeK Rossii) «Ob utverzhdenii Sostava i soderzhanija organizacionnyh i tehnicheskih mer po obespecheniju bezopasnosti personal'nyh dannyh pri ih obrabotke v informacionnyh sistemah personal'nyh dannyh»* [Order of the Federal Service for Technical and Export Control (FSTEC Russia) "On approval of the composition and content of organizational and technical measures to ensure the security of personal data at their processing in information systems of personal data"]. 2013. no 21. (In Russ.).
21. *Prikaz FSTJeK Rossii «Ob utverzhdenii sostava i soderzhanija organizacionnyh i tehnicheskih mer po obespecheniju bezopasnosti personal'nyh dannyh pri ih obrabotke v informacionnyh sistemah personal'nyh dannyh».* [Order FSTEC Russia "On approval of the composition and content of organizational and technical measures to ensure the security of personal data at their processing in information systems of personal data"]. 11.02.2013. no 17. (In Russ.).
22. *Polozhenie Central'nogo Banka Rossijskoj Federacii «O trebovanijah k obespecheniju zashhity informacii pri osushhestvlenii perevodov denezhnyh sredstv i o porjadke osushhestvlenija Bankom Rossii kontrolja za sobljudeniem trebovanij k obespecheniju*

zashhity informacii pri osushhestvlenii perevodov denezhnyh sredstv». [The position of the Central Bank of the Russian Federation "On requirements to ensure the protection of sensitive information in the transfer of funds and the exercise of the Bank of Russia control over compliance with requirements to ensure the protection of sensitive information in the transfer of funds"]. 2012. no 382-P. (In Russ.).

23. *Postanovlenie Pravitel'stva Rossijskoj Federacii «Ob utverzhdenii Polozhenija o zashhite informacii v platezhnoj sisteme»* [Resolution of the Government of the Russian Federation "On Approval of the Regulations on the Protection of the information in the payment system"]. 2012. no. 584. (In Russ.).
24. Payment Card Industry Data Security Standard (PCI DSS). PCI Security Standards Council LLC. Version 2.0. 2010. 75 p. URL: https://www.pcisecuritystandards.org/documents.
25. *Standart Banka Rossii STO BR IBBS-1.0-2014 Obespechenie informacionnoj bezopasnosti organizacij bankovskoj sistemy Rossijskoj Federacii. Obshhie polozhenija.* [The Standard Bank of Russia STO BR IBBS-1.0-2014 Information security organizations of the banking system of the Russian Federation. General provisions. - Instead of STO BR IBBSS-1.0-2010; Introduced 06/01/2014]. M.: Vestnik Banka Rossii. 2014. no. 48–49. 37 p. (In Russ.).
26. GOST R ISO/MJeK 27001—2006. [Information technology. Methods and means of ensuring safety. Information security management systems. Requirements.]. M.: Standartinform. 2008. 26 p. (In Russ.).
27. GOST R ISO/MJeK 13335-1- 2006. [Information technology. Methods and means of ensuring safety. Part 1: Concepts and models for the management of security of information and telecommunication technologies.]. M.: Standartinform. 2006. 23 p. (In Russ.).
28. Osipov V. Ju., Nosal' I. A. [Substantiation of the period of revision of information security measures]. *Informacionno–upravljajushhie sistemy - Information and Control Systems.* 2014. no. 1. pp. 63–69. (In Russ.).
29. Osipov V. Ju., Nosal' I. A. [Substantiation of information security measures]. *Informacionno–upravljajushhie sistemy – Information and Control Systems.* 2013. no. 2. pp. 48–53. (In Russ.).
30. Nosal' I. A. [Justification of the optimal set of permissions]. *Kompleksnaja zashhita ob"ektov informatizacii i izmeritel'nye tehnologii: Sb. nauchn. tr. Vserossijskoj nauchno-prakticheskoj konf. s mezhdunar. uchast.* [Comprehensive protection of information objects and measurement technology: Sat. Scien. tr. All-Russian Scientific-Practical Conference. with int. participation]. Sankt-Peterburg:Izdatel'stvo Politehnicheskogo universiteta. 2014. pp. 41–45.

Носаль Ирина Алексеевна — аспирант лаборатории прикладной информатики и проблем информатизации общества, Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН). Область научных интересов: защита информации, информационная безопасность. Число научных публикаций — 6. ironia.i@gmail.com; 14-я линия В.О., д. 39, Санкт-Петербург, 199178; р.т.: +7(812)3281113, Факс: +7 (812)3284450.

Nosal Irina Alexeevna — Ph.D. student of laboratory of Applied Informatics and Problems of Society Informatization, St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences (SPIIRAS). Research interests: information security. The number of publications — 6. ironia.i@gmail.com; 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone: +7 (812)328-1113, Fax: +7 (812)3284450.

РЕФЕРАТ

Носаль И.А. **Метод обоснования мероприятий информационной безопасности социально-важных объектов.**

Исследуется процесс обеспечения информационной безопасности в структурах Пенсионного фонда Российской Федерации, как социально-важных объектов. Отмечается, что существующее научно-методическое обеспечение такой безопасности не совершенно. Для повышения информационной безопасности в структурах Пенсионного фонда Российской Федерации предлагается усовершенствованный метод поиска целесообразных мероприятий защиты от возможных угроз. Разработана совокупность типовых моделей для оценки информационной безопасности этих объектов, применимых в рамках предложенного метода. Разработанные метод и модели ориентированы на обоснование мероприятий информационной безопасности в изменяющихся внутренних и внешних условиях. Приведены результаты моделирования, подтверждающие справедливость предложенного метода и входящих в него моделей.

SUMMARY

Nosal I.A. **Method of Information Security Measures Substantiation for Socially Important Objects.**

The process of information security in the structures of the Pension Fund of the Russian Federation, as a socially important objects, is studied. It is noted that the existing scientific and methodological support of such security is not perfect. To improve information security in the structures of the Pension Fund of the Russian Federation an improved method of searching for appropriate measures to protect against potential threats is proposed. The set of standard models for the evaluation of information security of these objects that are applicable within the proposed method are developed. The methods and models are focused on study of information security measures in the changing internal and external conditions. The simulation results confirm the validity of the proposed method and its constituent models.

УДК 004.045; 004.773

А.М. РОМАНЧЕНКО

МЕТОД ОЦЕНИВАНИЯ РЕЗУЛЬТАТОВ КРИПТОАНАЛИЗА БЛОЧНОГО ШИФРА

Романченко А.М. **Метод оценивания результатов криптоанализа блочного шифра.**
Аннотация. Данная работа направлена на разработку метода численного оценивания криптостойкости блочного шифра к различным методам криптоанализа при заданных ограничениях. Его использование позволяет сравнивать криптостойкость разных шифров и быстро определять возможность их взлома на практике.
Ключевые слова: блочные шифры, криптоанализ, оценивание криптостойкости.

Romanchenko A.M. **The method of Evaluation of the Results of a Block Cipher Cryptanalysis.**
Abstract. This work aims to develop a method of numerical estimation of the reliability block cipher cryptanalysis to various methods under given constraints. Its use allows you to compare different cryptographic ciphers and quickly determine the possibility of breaking into practice.
Keywords: block ciphers, cryptanalysis, evaluation of cryptographic ciphers.

1. Введение. Криптографические алгоритмы являются неотъемлемой частью информационно-телекоммуникационных систем, в том числе и специального назначения. При этом анализ уязвимостей таких систем связанных с использованием криптографических алгоритмов - это важнейшая составляющая безопасности в информационной сфере. Любая уязвимость, связанная с криптографическими алгоритмами, может стать важным преимуществом для одной из сторон и потенциально способна привести к утечке информации ограниченного распространения.

Все уязвимости, связанные с использованием криптографических алгоритмов, можно разделить на несколько классов. Этими классами являются:

– уязвимости связанные с криптостойкостью используемых алгоритмов шифрования;

– уязвимости связанные с некорректным использованием(реализацией) алгоритмов;

– уязвимости криптографических сетевых протоколов.

Первые два класса уязвимостей реализуются на практике путем проведения криптоанализа используемых алгоритмов, а третий, кроме криптоанализа, включает в себя контроль среды передачи данных и оказание активных воздействий на информационную систему.

Одними из наиболее значимых являются уязвимости связанные с криптостойкостью блочных шифров. Исходными данными проведения криптоанализа блочного алгоритма шифрования является его спецификация. Зная алгоритм шифрования, данные для проверки

работоспособности модели можно сгенерировать самостоятельно. Результатом криптоанализа являются частные модели алгоритма шифрования (математические, вероятностные и т. п.), соответствующие конкретным методам и их параметры, на основе которых можно сделать вывод о реализуемости этого метода криптоанализа на практике. Наиболее существенными параметрами модели является объем и вид исходных данных, а также объем вычислительных ресурсов, требуемых для проведения криптоанализа. Практическую значимость произвольного метода криптоанализа с точки зрения объема исходных данных и объема вычислений можно оценить с помощью графика см. рисунок 1.

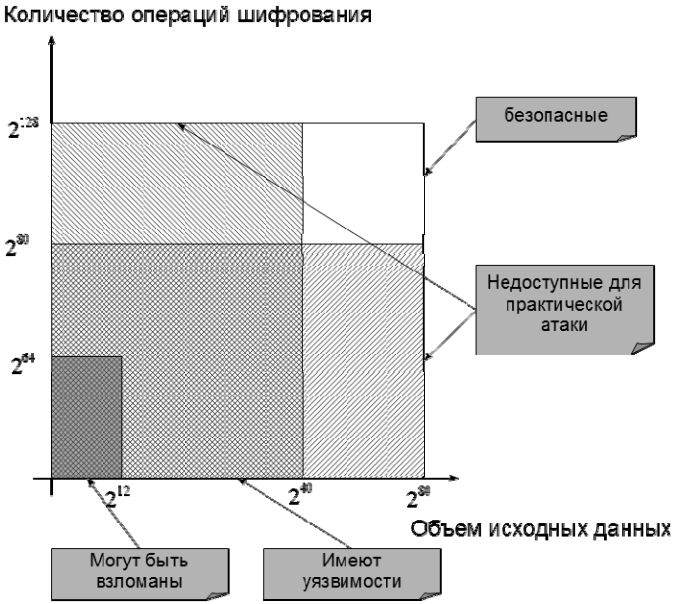


Рис. 1. Оценивание практической применимости различных методов криптоанализа

В данной работе предлагается комплексный подход к оцениванию криптостойкости блочных шифров с использованием обобщенного показателя криптостойкости. Преимуществом этого подхода является отражение криптостойкости алгоритма к нескольким методам криптоанализа с помощью одного числового показателя. Это дает возможность сравнительного анализа характеристик различных алгоритмов и быстрого оценивания их криптостойкости.

2. Оценивание результатов криптоанализа блочных шифров с использованием обобщенного показателя криптостойкости. Предлагается при оценивании криптостойкости произвольного блочного шифра использовать следующие методы криптоанализа:

– дифференциальный криптоанализ [4,7] и его разновидности если они дают результат лучше чем классический дифференциальный криптоанализ;

– линейный криптоанализ [5,6,7] и его разновидности если они дают результат лучше чем классический линейный криптоанализ;

– интегральный криптоанализ [8];

– алгебраический криптоанализ [9,10];

– метод полного перебора.

Предлагается в общем случае не использовать в составе обобщенного показателя криптостойкости результаты технических методов криптоанализа, так как в современных условиях эту атаку практически невозможно провести на практике – аппаратура шифрования, как правило, удалена от местонахождения криптоаналитика и какое-либо воздействие на нее не представляется возможным.

Результаты проведения криптоанализа блочных алгоритмов шифрования целесообразно представлять в форме обобщенного показателя криптостойкости. Обобщенный показатель стойкости формируется следующим образом:

$$F_{\Sigma} = \sum_{i=1}^{s} f_i(W, N, P_d)\,;\; F_{\Sigma} < F_d,$$

где f_i является функцией, характеризующей стойкость БШ к одному виду криптоанализа при заданных ограничениях количества исходных тестов W, количества вычислительных ресурсов N (выраженного в числе операций шифрования/расшифрования), и заданной вероятности вскрытия алгоритма шифрования при этих ограничениях. Количество методов криптоанализа, использованных в методике, обозначено как s. Функция f_i вычисляется как:

$$f_i(W, N, P_d) = \begin{cases} 0, \forall\, W, N\ P_{W,N} < P_d \\ 1, P_{W,N} \geq P_d \\ \dfrac{P_d}{\left(1 + \dfrac{\acute{W}}{W}\right)\left(1 + \dfrac{\acute{N}}{N}\right)}, P_{W,N} < P_d, \exists \acute{W}, \acute{N}\colon P_{\acute{W}\acute{N}} \geq P_d \end{cases},$$

т.е. она принимает значение 0 если при любых ограничениях вероятность успешной атаки меньше заданной вероятности P_d,

значение 1 если при заданных ограничениях вероятность успешной атаки больше или равна заданной вероятности. Если существуют такие условия, при которых вероятность успешной атаки больше или равна заданной вероятности, то значение функции вычисляется как отношение заданной вероятности к отношению количества требуемых исходных данных и имеющихся исходных данных, и отношению требуемых и имеющихся вычислительных ресурсов.

Для обобщенного показателя стойкости показателя была выбрана аддитивная форма, так как величина вклада отдельных составляющих, соответствующих определенному методу криптоанализа, не зависит от вкладов других методов. Превышение данным показателем порога $F_d = s * 10^{-6}$ будет означать, что исследуемый БШ является потенциально уязвимым по крайней мере для одной криптоаналитической атаки. Физический смысл данного показателя состоит в том, что он отражает существование хотя бы одного практического метода криптоанализа при значении показателя близком к 1, при значении показателя от 1 до N существует несколько практических методов криптоанализа, и при значении намного меньшем 1 он показывает, какое минимальное количество ресурсов (исходных данных или вычислительных мощностей) необходимо добавить к уже имеющимся для проведения хотя бы одной практической атаки.

При таком подходе к вычислению стойкости БШ по значению этого обобщенного показателя сразу можно определить, является ли алгоритм шифрования устойчивым ко всем видам криптоаналитических атак или нет. Если значение показателя близко к 1, то целесообразно проведение одной из криптоаналитических атак, если значение показателя равно 0, то надежность алгоритма не позволяет его вскрыть даже теоретически, без существенных изменений в теории криптоанализа, и если значение показателя значительно меньше 1 то целесообразно проведение дальнейших исследований в направлении усовершенствования атаки или смягчения ограничений на применение атаки.

3. Пример реализации предложенного подхода к анализу криптостойкости блочного шифра. Рассмотрим пример анализа криптостойкости для двух практических блочных шифров — алгоритма шифрования DES и блочного шифра ГОСТ 28147-89.

Алгоритм DES в настоящее время используется ограниченно, так как он имеет недостаточную длину ключа, что позволяет вскрыть его методом полного перебора. Несмотря на то, что реализация криптоанализа методом полного перебора недоступна для одиночных

пользователей и небольших компаний из-за высокой стоимости вычислительных ресурсов, крупным компаниям и государственным службам такая реализация метода криптоанализа доступна.

Блочный шифр ГОСТ 28147-89, напротив, в настоящее время является государственным стандартом шифрования и повсеместно применяется для шифрования большого объема данных.

В качестве исходных данных предположим, что в наличии имеется $W = 2^{30}$ открытых текстов и соответствующих им шифртекстов, зашифрованных на одном ключе, что будет иметь место при шифровании и передаче стандартного фильма в одном сеансе связи. Объем вычислительных ресурсов имеющихся в наличии примем за 1024 стандартных машины класса Pentium 4 с частотой 3 ГГц, что соответствует потенциальным ресурсам небольшой организации или хакерской группы. При шифровании примерно со скоростью 150 Мбайт в секунду, которая близка к теоретическому пределу с учетом 4-х ядер и оптимизированной реализации, это составит порядка 2^{24} в секунду для одной машины и 2^{34} в секунду для 1024 машин. Время отводимое на расшифрование возьмем 8 суток, то есть $60 * 60 * 24 * 8 = 2^{20}$ за которое можно выполнить 2^{54} операций шифрования. Результаты отдельных видов криптоанализа выбранных алгоритмов по материалам открытых публикаций([1]-[5]) приведены в таблице 1.

Таблица 1. Результаты криптоанализа выбранных блочных шифров по материалам открытых публикаций

	DES	ГОСТ 28147-89
Линейный криптоанализ	$W = 2^{43}, N = 2^{43}$	-
Дифференциальный криптоанализ	$W = 2^{55}, N = 2^{55}$	-
Алгебраический криптоанализ	-	$W = 2^{64}, N = 2^{248}$
Интегральный криптоанализ	-	-
Полный перебор	$W = 1, N = 2^{56}$	$W = 1, N = 2^{256}$

$$F_{\Sigma}(DES) = \frac{1}{\left(1 + \frac{2^{43}}{2^{30}}\right) * 1} + \frac{1}{\left(1 + \frac{2^{55}}{2^{30}}\right) * \left(1 + \frac{2^{55}}{2^{54}}\right)} + \frac{1}{1 * \left(1 + \frac{2^{56}}{2^{54}}\right)}.$$

$$F_{\Sigma}(DES) \approx 1.2e - 4 + 9.9e - 9 + 0.2 \approx 0.200122.$$

Величина обобщенного показателя криптостойкости для алгоритма DES говорит о том, что при заданных условиях его вскрытие невозможно, но при некотором ослаблении ограничений его взлом является реальностью. То есть признаком потенциальной слабости является близость порядка этого показателя к 1. Он отражает

реальную возможность взлома данного алгоритма методом полного перебора, но с несколько большей вычислительной сложностью, чем выбрана нами в условиях задачи.

$$F_{\Sigma}(GOST) = \frac{1}{\left(1+\frac{2^{64}}{2^{30}}\right) * \left(1+\frac{2^{248}}{2^{54}}\right)} + \frac{1}{1 * \left(1+\frac{2^{256}}{2^{54}}\right)}.$$

$$F_{\Sigma}(GOST) \approx 2.31e-69 + 1.5e-61 \approx 1.555e-61.$$

Для алгоритма ГОСТ очевидно, что порядок показателя криптостойкости очень далек далек от единицы, и это говорит о том, что даже с учетом существенных послаблений в ограничениях взлом этого алгоритма не представляется возможным, что и соответствует реальному положению дел - на сегодняшний день алгоритм ГОСТ 28147-89 не имеет значимых уязвимостей [2, 3].

4. Заключение. В данной работе предложен метод оценивания результатов криптоанализа блочных шифров. Он позволяет получить численную характеристику криптостойкости блочного шифра к нескольким методам криптоанализа. Метод основан на использовании обобщенного показателя криптостойкости который учитывает вычислительные ресурсы и объем исходных данных имеющиеся в распоряжении криптоаналитика. Данный метод может быть использован при определении практической возможности взлома различных блочных шифров и сравнения их между собой по этому показателю.

Литература

1. *Панасенко С.П.* Стандарт шифрования ГОСТ 28147-89. Обзор криптоаналитических исследований. URL: http://www.inssl.com/standart-of-cipher.html (дата обращения: 23.03.2015).
2. *Shorin V.V., Jelezniakov V.V., Gabidulin E.M.* Linear and Differential Cryptanalysis of Russian GOST // Electronic Notes in Discrete Mathematics. 2001. vol. 6. pp 538–547.
3. *Courtois N.* Security Evaluation of GOST 28147-89 In View Of International Standardisation // Cryptologia. 2012. vol. 36(1). pp. 2–13.
4. *Biham E., Shamir A.* Differential Cryptanalysis of the Data Encryption Standard // Springer-Verlag Computers. 1993. 188 p.
5. *Matsui M.* Linear cryptanalysis method for DES cipher // In Advances in Cryptology - EUROCRYPT'93. Springer-Verlag. 1993. LNCS 765. pp. 386–397.
6. *Biham E.* On Matsui Linear Cryptoanalysis // In Advances in Cryptology – EUROCRYPT '94. Springer-Verlag. 1995. LNCS 950. pp. 341–355.
7. *Keliher L.* Refined analysis of bounds related to linear and differential cryptanalysis for the AES // Fourth Conference on the Advanced Encryption Standard (AES4). Springer-Verlag. 2005. LNCS 3373. pp. 42–57.

8. *Knudsen L., Wagner D.* Integral cryptanalysis // Fast Software Encryption. Springer-Verlag. 2002. LNCS 2365. pp. 112–127.
9. *Courtois N.T., Pieprzyk J.* Cryptanalysis of Block Ciphers with Overdefined Systems of Equation //In Proceeding of Asiacrypt 2002. Springer-Verlag. 2002. LNCS 2501. pp. 378–385.
10. *Biryukov A., De Canniere C.* Block Ciphers and Systems of Quardatic Equations // In Fast Software Encryption. Springer-Verlag. 2003. LNCS 2887. pp. 274–289.

References

1. Panasenko S.P. GOST 28147-89 encryption standard. [Overview of cryptanalytic research]. Available at http://www.inssl.com/standart-of-cipher.html. (accessed: 23.03.2015). (In Russ.)
2. Shorin V.V. Jelezniakov V.V., Gabidulin E.M. Linear and Differential Cryptanalysis of Russian GOST. *Electronic Notes in Discrete Mathematics*. 2001. vol. 6. pp 538–547.
3. Courtois N. Security Evaluation of GOST 28147-89 In View Of International Standardisation. *Cryptologia*. 2012. vol. 36(1). pp. 2–13.
4. Biham E., Shamir A. Differential Cryptanalysis of the Data Encryption Standard. Springer-Verlag Computers. 1993. 188 p.
5. Matsui M. Linear cryptanalysis method for DES cipher. In Advances in Cryptology - EUROCRYPT'93. Springer-Verlag. 1993. LNCS 765. pp. 386–397.
6. Biham E. On Matsui Linear Cryptoanalysis. In Advances in Cryptology – EUROCRYPT '94. Springer-Verlag. 1995. LNCS 950. pp. 341–355.
7. Keliher L. Refined analysis of bounds related to linear and differential cryptanalysis for the AES. Fourth Conference on the Advanced Encryption Standard (AES4). Springer-Verlag. 2005. LNCS 3373. pp. 42–57.
8. Knudsen L., Wagner D. Integral cryptanalysis. Fast Software Encryption. Springer-Verlag. 2002. LNCS 2365. pp. 112–127.
9. Courtois N.T., Pieprzyk J. Cryptanalysis of Block Ciphers with Overdefined Systems of Equation. In Proceeding of Asiacrypt 2002. Springer-Verlag. 2002. LNCS 2501. pp. 378–385.
10. Biryukov A., De Canniere C. Block Ciphers and Systems of Quardatic Equations. In Fast Software Encryption 2003. Springer-Verlag. 2003. LNCS 2887. pp. 274–289.

Романченко Александр Михайлович — к-т техн. наук, старший преподаватель кафедры систем сбора и обработки информации Военно-космическая академия имени А.Ф. Можайского. Область научных интересов: криптография, информационная безопасность, сетевые технологии. Число научных публикаций — 15. rcrst@newmail.ru; ул. Ждановская, д.13, Санкт-Петербург, 197198, РФ; р.т.: +7(812)237-19-60.

Romanchenko Alexander Mikhailovitch — Ph.D., senior lecturer of the information acquisition and data processing department, Mozhaisky Military Aerospace Academy. Research interests: cryptography, information security, network technology. The number of publications — 15. rcrst@newmail.ru; Zdanovskaya str.13, Saint-Petersburg, Russia, 197198; office phone: +7(812)237-19-60.

РЕФЕРАТ

Романченко А.М. **Метод оценивания результатов криптоанализа блочного шифра.**

Данная работа посвящена разработке обобщенного показателя криптостойкости который характеризует устойчивость блочного шифра к различным методам криптоанализа. Этот показатель формируется с учетом реальных возможностей криптоаналитика и учитывает количество исходных данных и вычислительных ресурсов используемых для анализа. Метод оценивания результатов криптоанализа основанный на использовании данного показателя позволит проводить сравнительный анализ криптостойкости разных шифров. Это показатель может быть использован, в том числе и для неизвестных сегодня методов, которые возможно будут открыты в будущем. При установке граничных значений предложенного показателя он может быть использован для формирования требований к криптостойкости алгоритмов для использования в аппаратуре или информационной системе. В работе приведены два примера использования метода для широко известных блочных алгоритмов шифрования – DES и ГОСТ 28147-89. Результаты применения метода совпадают с традиционными результатами оценивания криптостойкости этих алгоритмов шифрования.

SUMMARY

Romanchenko A.M. **The Method of Evaluation of the Results of a Block Cipher Cryptanalysis.**

This work is devoted to the development of a generalized indicator that characterizes the resistance of cryptographic block cipher to various methods of cryptanalysis. It is derived taking into account the real possibilities cryptanalyst and monitors the amount of input data and computing resources used for the analysis. The method of evaluation of the results of cryptanalysis based on the use of this indicator will allow for a comparative analysis of different cryptographic ciphers. This indicator may be used including methods for unknown today which might be visible in the future. When setting boundary values proposed indicator, it may be used for establishing the requirements for cryptographic algorithms for use in the apparatus or system information. The paper presents two examples of using the method for well-known block encryption algorithms - DES and GOST 28147-89. The results of applying the method coincide with the traditional evaluation of the reliability of the results of encryption algorithms.

УДК 621.3.089.2

О.О. Басов

ПРИНЦИПЫ ПОСТРОЕНИЯ ПОЛИМОДАЛЬНЫХ ИНФОКОММУНИКАЦИОННЫХ СИСТЕМ НА ОСНОВЕ МНОГОМОДАЛЬНЫХ АРХИТЕКТУР АБОНЕНТСКИХ ТЕРМИНАЛОВ

Басов О.О. **Принципы построения полимодальных инфокоммуникационных систем на основе многомодальных архитектур абонентских терминалов.**
Аннотация. Анализ существующих многомодальных интерфейсов, их основных характеристик и областей применения, а также результатов общих исследований в области многомодального взаимодействия и дизайна интерфейсов позволил сделать вывод о возможности построения полимодальных инфокоммуникационных систем на основе многомодальных архитектур их абонентских терминалов. Для решения задач межличностной коммуникации через технические средства связи в работе предлагаются принципы построения полимодальных систем и иерархическая система их моделей.
Ключевые слова: полимодальная инфокоммуникационная система, принципы построения, иерархическая система моделей, задача синтеза, макромодель.

Basov O.O. **Principles of Construction of Polymodal Info-Communication Systems based on Multimodal Architectures of Subscriber's Terminals.**
Abstract. Analysis of the existing polymodal interfaces, their main characteristics and applications, as well as the results of common investigations in the field of multimodal interaction and interface design led to make a conclusion about the possibility of building a polymodal infocommunication systems based on multimodal architectures of subscriber's terminals. To solve the tasks of interpersonal communication through technical means of communication the principles of polymodal systems construction and hierarchical system of their models are suggested in the article.
Keywords: polymodal infocommunication system, principles of construction, hierarchical system of models, synthesis task, the macro model.

1. Введение. В [1] было показано, что при разработке полимодальной инфокоммуникационной системы (ПИКС) ее удобно представлять в виде совокупности абонентских терминалов (рисунок 1) и сети передачи данных (СПД). При ориентации на предоставление единственной услуги в виде «соединение с сетью» практически важными являются две основные научно-методические задачи синтеза ПИКС.

Прямая. Имеются сведения об объемах сообщений (комплексе модальностей), которые необходимо передать. Требуется определить объем ресурсов СПД, чтобы обеспечить требуемое качество приема переданной информации.

Двойственная. Имеются заданные ресурсы СПД. Требуется передать максимальный объем сообщений заданного качества, в том числе с оптимальным выбором модальностей и методов их обработки (объединения/разделения, синхронизации, кодирования).

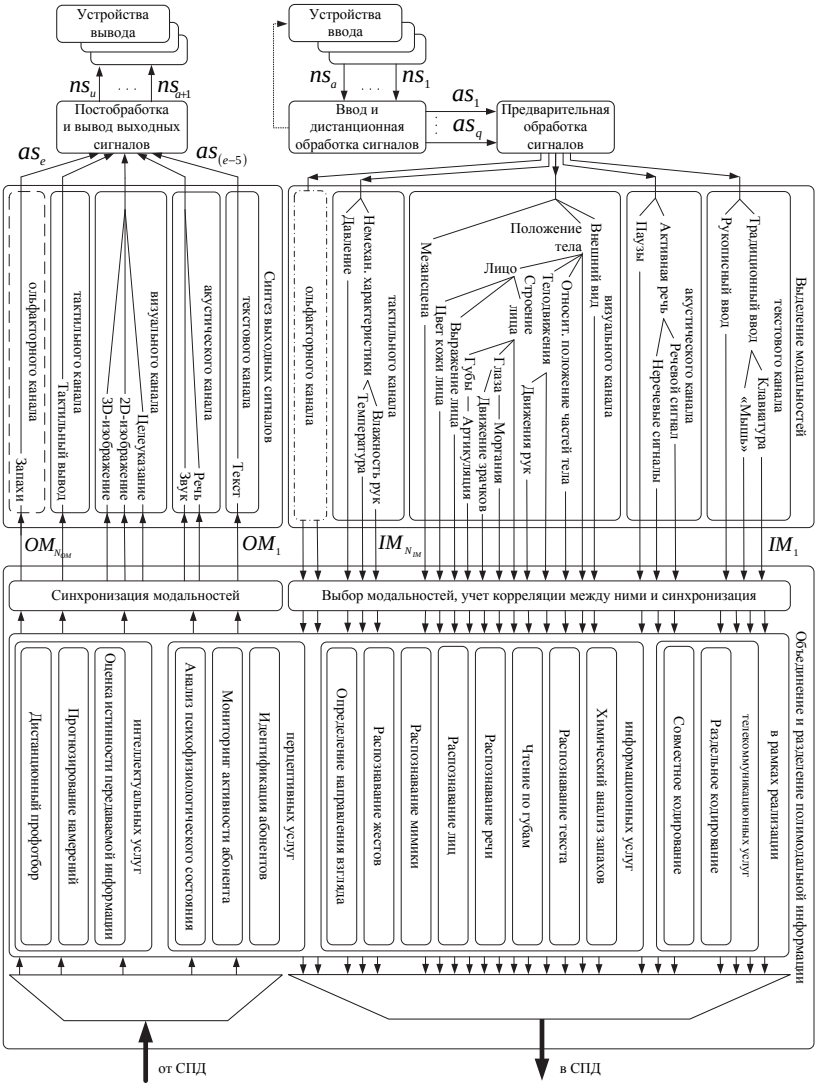


Рис. 1. Обобщенная структурная схема (архитектура) абонентского терминала полимодальной инфокоммуникационной системы

Понятно, что если бы речь шла только о СПД и услугах связи, то опыт решения таких задач в предметной области накоплен доста-

точный. При синтезе ПИКС необходима разработка новых теоретических конструкций, учитывающих необходимость обеспечения взаимодействия процедур обработки модальностей и алгоритмов передачи соответствующих им блоков данных через СПД.

2. Принципы построения ПИКС. В результате исследования установлено, что построение ПИКС должно базироваться на *основополагающих принципах*, которые можно разделить на следующие группы.

I. Общеметодологические.

1. *Принцип соответствия* состояния ПИКС ситуации в системе управления. Для его выполнения при синтезе ПИКС необходимо обеспечить нахождение для каждой ситуации в системе управления соответствующего состояния инфокоммуникационной системы (требуемых комбинации модальностей и модели реализуемой услуги), оптимального с точки зрения целостности полимодальной информации.

2. *Принцип системности* подразумевает взаимообусловленность полимодальности абонентских терминалов с характеристиками обеспечивающих подсистем ПИКС во всем спектре идентифицируемых условий коммуникативного взаимодействия.

3. *Принцип функциональной полноты* требует размещения в абонентских терминалах ПИКС всех требуемых типов устройств ввода/вывода, функциональных (программных, аппаратных или программно-аппаратных) модулей обработки сигналов и реализации услуг, а также соответствующих обеспечивающих подсистем.

4. *Принцип мультисервисности* [2], под которой понимается независимость технологий предоставления услуг от транспортных технологий.

5. *Принцип полимодельности* подразумевает комплексное использование моделей входных и выходных сигналов различных модальностей, кодирования и передачи полимодальной информации и полимодальных (информационных, перцептивных и интеллектуальных) услуг.

II. Методические.

6. *Принцип открытости архитектуры* состоит в возможности расширения числа идентифицируемых в системе управления ситуаций с соответствующей доработкой абонентского терминала и транспортной инфраструктуры (СПД) при внедрении новых технологий и услуг.

7. *Принцип функциональной замкнутости* состоит в реализации полимодальности либо в системе в целом, либо в ее подсистеме, либо в отдельном функциональном элементе.

8. *Принцип оперативной управляемости.* Возможность переключения нагрузки (потока блоков данных) с выходов абонентских терминалов на входы оконечного оборудования СПД определяется оперативностью актуализации и полнотой описания ситуации в системе управления.

III. Прикладные.

9. *Принцип дифференцированности услуг и имеющихся ресурсов* заключается в предварительных процедурах определения для каждой услуги величины единицы канального ресурса [1], классификации потоков блоков данных по уровню требований (числу единиц канального ресурса, приоритетам, важности и пр.) и имеющихся ресурсов по соответствию этим потокам (числу единиц канального ресурса, удельной себестоимости сети). Выполнение этого принципа ограничивает пространство управления ПИКС, дискретезирует область поиска.

10. *Принцип децентрализации предоставляемых услуг (инвариантности доступа)* заключается в их независимом функционировании, при котором отключение или перемещение одного из физических модулей не влияет на работу системы в целом. Полная децентрализация услуг в ПИКС и свободная композиция их независимо от специфики функционирования системы управления возможна лишь при условии, когда каждая услуга не зависит от используемой технологии и способна расширять свои знания о предметной области, используя знания других услуг.

11. *Принцип ассоциативности и толерантности обращения к информации* позволяет быстро получить нужную информацию, независимо от объемов выборки. Субъективные воздействия на информацию необходимо снижать за счет возможности отмены и повтора действий, а также путем анализа различных форматов ввода и интерпретации любых разумных действий абонента.

12. *Принцип гарантированного доступа к контексту.* Гарантированный доступ к контексту должен быть обеспечен с помощью транспортной инфраструктуры ПИКС или самими источниками контекста, таким образом, чтобы каждый абонент имел доступ к контекстам других абонентов независимо от степени своей нагрузки или физической доступности. Без систематизации общего контекста коммуникативного акта невозможно формализованное представление контекста для всех его участников. Реализация контекстно-ориентированных услуг обязывает использовать базовую модель контекста предметной области коммуникативного акта, при этом пропадает необходимость включения в соответствующее приложение меха-

низмов поиска соответствий понятий контекстов различных источников.

IV. Практические.

13. *Принцип многооператорности* определяет возможность участия нескольких операторов в процессе предоставления услуги и разделение их ответственности в соответствии с их областью деятельности.

14. *Принцип обратной связи* предписывает сообщать абонентам о действиях ПИКС, ее подсистем и элементов, их реакциях, изменениях состояния или ситуации, об ошибках и исключениях. Сообщения должны быть четкими, краткими, однозначными и написанными на языке, понятном абоненту.

Реализация перечисленных принципов позволяет строить инфокоммуникационные системы вне рамок традиционных принципов разделения передаваемой при межличностной коммуникации информации на услуги. Для количественного обоснования соответствующих решений была разработана специальная система моделей ПИКС, ее подсистем и элементов.

3. Структура системы моделей полимодальных инфокоммуникационных систем. Базируясь на опыте моделирования, накопленном в предметной области [3–10], в основу сформированной системы моделей (рисунок 2) положены следующие принципы:

– *единства*, состоящий в представлении ПИКС любого типа в виде однообразной системы моделей основных функциональных модулей и обеспечивающих подсистем;

– *глобальности*, предписывающий производить формализацию ПИКС в целом, без его искусственной априорной декомпозиции на отдельные элементы (подсистемы) и их отдельного моделирования;

– *иерархичности*, подразумевающий использование моделей одного и того же объекта, различающихся уровнем представления (мета-, макро- и субуровня), степенью детализации исходных данных, целевыми функциями и составом системы ограничений;

– *многоцелевой ориентации*, позволяющий использовать систему моделей в проектных ситуациях, различающихся степенью детализации исходных данных, целевыми функциями и составом системы ограничений;

Рис. 2. Структура и состав системы моделей ПИКС

аналитичности, заключающийся в формульном представлении моделей макро- и субуровней системы;

упрощения, основанный на исключении несущественных (на данном уровне исследований) факторов, объединении, линеаризации и регуляризации менее существенных переменных;

надежности вычислений, требующий обеспечения устойчивой работы алгоритмов поиска как в допустимой, так и недопустимой областях функциональных характеристик, что позволяет разрешить проблему выбора начальной точки задач комбинаторной оптимизации.

Описательные модели (рисунок 2) предназначены для обобщения, систематизации и классификации положений исходного методологического базиса. Они способствуют четкому ограничению области применения теоретических и методологических конструкций, а также предлагаемых ниже моделей мета-, макро- и субуровней.

Обобщенная модель ПИКС является теоретической конструкцией метауровня, связывая (принцип дополнительности Бора) его через макромодель ПИКС с моделями субуровня исследований, проводимых в предметной области.

Блок теоретических конструкций объединяет архитектуры, схемы, формализованные правила и алгоритмы отношений, объективно существующие в предметной области, в том числе, реализованные в соответствующих абонентских терминалах и СПД.

Математические модели предназначены для формального (посредством математических средств) описания зависимости внешних функциональных характеристик исследуемых объектов (ПИКС, подсистем, компонентов) от их внутренних характеристик и имеющихся исходных данных.

Макромодель представляет собой теоретико-множественную модель, объединяющую математические и аналитико-алгоритмические страты, оперирующие макрохарактеристиками инфраструктуры ПИКС (рисунок 3).

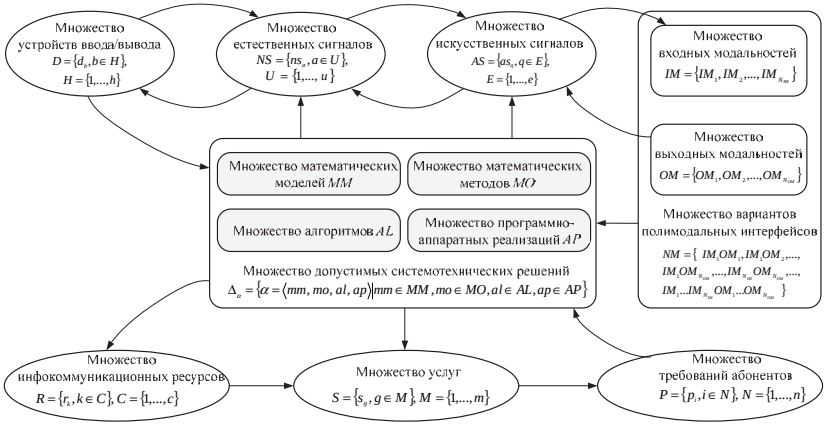


Рис. 3. Теоретико-множественная модель ПИКС

К основным из них относятся [10, 11]:

– множество требований $P=\{p_i, i\in N\}$, $N=\{1,...,n\}$ пользователей к ПИКС, на удовлетворение которых направлено множество услуг $S=\{s_g, g\in M\}$, $M=\{1,...,m\}$, реализуемых с использованием инфокоммуникационных ресурсов $R=\{r_k, k\in C\}$, $C=\{1,...,c\}$;

– множество устройств ввода/вывода сигналов различных модальностей, доступных абоненту $D=\{d_b, b\in H\}$, $H=\{1,...,h\}$;

– множество преобразований $W=\{w_f, f\in O\}$, $O=\{1,...,o\}$, выполняемых при реализации услуги;

– множества потоков искусственных $AS = \{as_q, q \in E\}$, $E = \{1,...,e\}$ и естественных сигналов $NS = \{ns_a, a \in U\}, U = \{1,...,u\}$, использующихся для анализа входных $IM = \{IM_1, IM_2,..., IM_{N_{IM}}\}$ и синтеза выходных $OM = \{OM_1, OM_2,..., OM_{N_{OM}}\}$ модальностей;

– множество вариантов многомодальных интерфейсов абонентских терминалов $NM = \{IM_1OM_1, IM_1OM_2,..., IM_1OM_{N_{OM}},...,$ $...IM_{N_{IM}}OM_{N_{OM}},..., IM_1...IM_{N_{IM}}OM_1...OM_{N_{OM}}\}$, формируемое за счет комбинации входных и выходных модальностей;

– множество допустимых системотехнических решений Δ_α, включающее в себя множества математических моделей *MM*, методов *MO*, алгоритмов *AL* и программно-аппаратных реализаций *AP* услуг, представленное в следующем виде:

$$\Delta_\alpha = \{\alpha = \langle mm, mo, al, ap \rangle | mm \in MM, mo \in MO, al \in AL, ap \in AP\}.$$

Для обработки сигналов в режиме реального времени T вводится множество $W^{(\alpha)}: AS \times NS \times T \to AS^{(\alpha)} \times NS^{(\alpha)}$, ограничивающее множество преобразований W на множестве решений Δ_α.

На эффективность передачи полимодальной информации влияют следующие ограничения:

1) на способы человеко-машинного взаимодействия со стороны абонента, связанные с его навыками использования абонентским терминалом (интерфейсом), информационных технологий, личными предпочтениями и физическими ограничениями: $UC = \{UC_i, i \in X\}$;

2) на способы человеко-машинного взаимодействия со стороны абонентского терминала, связанные с его программно-аппаратными возможностями: $DC = \{DC_j, j \in Y\}$;

3) среды человеко-машинного взаимодействия (тип помещения и уровень шумов в нем, число абонентов, расстояние между абонентом и абонентским терминалом, и другие): $EC = \{EC_k, k \in Z\}$;

4) предоставляемых услуг, связанные с предметной областью, наличием доступа к инфокоммуникационным ресурсам, их объемом и типом: $SC = \{SC_l, l \in V\}$.

Для формирования множества Δ_α введены подмножества декартовых произведений исходных множеств макромодели, задающих пространство альтернатив синтеза:

$$F_{UC}^{(\alpha)} \subseteq P^{(\alpha)} \times S^{(\alpha)} \times R^{(\alpha)} \times D^{(\alpha)} \times AS^{(\alpha)} \times NS^{(\alpha)};$$
$$F_{DC}^{(\alpha)} \subseteq P^{(\alpha)} \times S^{(\alpha)} \times R^{(\alpha)} \times D^{(\alpha)} \times AS^{(\alpha)} \times NS^{(\alpha)};$$
$$F_{EC}^{(\alpha)} \subseteq P^{(\alpha)} \times S^{(\alpha)} \times R^{(\alpha)} \times D^{(\alpha)} \times AS^{(\alpha)} \times NS^{(\alpha)};$$
$$F_{SC}^{(\alpha)} \subseteq P^{(\alpha)} \times S^{(\alpha)} \times R^{(\alpha)} \times D^{(\alpha)} \times AS^{(\alpha)} \times NS^{(\alpha)}.$$

С учетом этого задача синтеза ПИКС сводится к поиску оптимальных (квазиоптимальных) подходов к формированию множества Δ_α с учетом ограничений UC, DC, EC, SC, реализующих систему на базе множества модальностей $NM^{(\alpha)}$:

$$\Delta_\alpha^{\text{огр}} = \begin{cases} \left\langle p_i^{(\alpha)}, s_g^{(\alpha)}, d_b^{(\alpha)}, r_k^{(\alpha)}, as_q^{(\alpha)}, ns_a^{(\alpha)} \right\rangle; \\ \Phi^{(\alpha)} : F_{UC}^{(\alpha)} \cap F_{DC}^{(\alpha)} \cap F_{EC}^{(\alpha)} \cap F_{SC}^{(\alpha)} \to B_m; \\ W^{(\alpha)} : AS \times NS \times T \to AS^{(\alpha)} \times NS^{(\alpha)}. \end{cases}$$

Выбор полной комбинации модальностей, допустимых в проектируемой системе, будет определяться следующим образом:

$$\overline{\Delta}_\alpha^{\text{огр}} = \left\{ \overline{\Theta}_\alpha(NM) \,\middle|\, \Psi^{(\alpha)} : \Theta_\alpha(NM) \times \Delta_\alpha^{\text{огр}} \to B_m \right\},$$

где $\Theta_\alpha(NM)$ – множество комбинаций модальностей, а элементы множества B_m принимают значения $\{0,1\}$.

На основе разрабатываемой теории необходим обоснованный выбор конкретных вариантов реализаций отображений $\Phi^{(\alpha)}$ и $\Psi^{(\alpha)}$, определяющих структуру и функции ПИКС и конфигурацию программно-аппаратного обеспечения, необходимого для ее реализации.

Целью макромоделирования является получение «экспресс-оценок» макрохарактеристик ПИКС для их дальнейшего автономного, но взаимосогласованного формального описания и ограничение области поиска оптимальной информационно-алгоритмической структуры ПИКС.

Математические модели $MM \in \Delta_\alpha$ являются *моделями субуровня* (рисунок 2) и предназначены для формализации зависимости внешних функциональных характеристик подсистем ПИКС от параметров их программно-аппаратных компонентов.

Модели кодирования и передачи полимодальной информации, обеспечивающие взаимное соответствие скорости обработки модальностей в абонентских терминалах (производительности источника ин-

формации) и производительности каналов связи и узлов коммутации в СПД, представлены в [12, 13]. На их основе реализуются коммуникативная и интерактивная стороны межличностного общения, усиленные передачей информации о его перцептивной стороне.

4. Заключение. Многомодальные интерфейсы абонентских терминалов ПИКС должны объединять входную информацию от множества различных сенсоров, пассивных и активных форм пользовательского ввода, иметь способность адаптироваться к пользователю, задаче, текущему диалогу (полилогу) и условиям среды функционирования. Полимодальные системы дадут абоненту широкие возможности по выражению своих действий и команд, а также лучшие средства для управления процессом визуализации мультимедийного вывода информации. Однако, организовать коммуникативное взаимодействие представляется возможным только в том случае, если абонентские терминалы ПИКС находятся в зонах взаимодействия с абонентами и связи с инфокоммуникационными ресурсами, а их многомодальные интерфейсы соответствуют физическим возможностям и предпочтениям абонента и могут обеспечить коммуникацию в текущих условиях среды функционирования. Для формализации и решения задачи построения (синтеза) ПИКС при данных ограничениях была предложена иерархическая система моделей.

В большинстве существующих приложений для получения информации пользователь вынужден идти на компромисс между естественностью взаимодействия и функциональными возможностями сервисов/устройств. В рамках предложенных моделей мета- и субуровней возможный набор естественных входных и выходных модальностей, а также моделей реализуемых полимодальных услуг определяется на этапе проектирования абонентского терминала ПИКС.

Отказ от традиционных принципов разделения передаваемой информации на услуги связи в пользу многомодального представления информации требует разработки моделей субуровня, учитывающих необходимость обеспечения взаимодействия процедур обработки модальностей и алгоритмов передачи блоков данных через СПД.

Литература

1. *Басов О.О., Саитов И.А.* Качество функционирования и эффективность полимодальных инфокоммуникационных систем // Труды СПИИРАН. 2014. Вып. 1(32). С. 152–170.
2. Концептуальные положения по построению мультисервисных сетей на ВСС РФ // М.: Минсвязи РФ. 2002.
3. *Ронжин А.Л., Карпов А.А., Ли И.В.* Речевой и многомодальные интерфейсы // М.: Наука. 2006. 173 с.

4. *Шелухин О.И., Лукьянцев Н.Ф.* Цифровая обработка и передача речи / под ред. О. И. Шелухина // М.: Радио и связь. 2000. 456 с.
5. *Кипяткова И.С., Ронжин А.Л., Карпов А.А.* Автоматическая обработка разговорной русской речи / монография // СПб.: ГУАП, 2013. 314 с.
6. *Гонсалес Р., Вудс Р.* Цифровая обработка изображений // М.: Техносфера. 2006. 1072 с.
7. *Устинов А.А.* Стохастическое кодирование видео- и речевой информации / монография: в 2 ч. / под ред. проф. В.Ф. Комаровича // СПб.: ВАС. 2005.
8. *Степанов С.Н.* Основы телетрафика мультисервисных сетей // М.: Эко-Трендз. 2010. 392 с.
9. *Саитов И.А.* Основы теории построения защищенных мультипротокольных оптических транспортных сетей телекоммуникационных систем / монография // Орел: Академия ФСО России. 2008. 220 с.
10. *Ронжин А.Л., Карпов А.А.* Проектирование интерактивных приложений с многомодальным интерфейсом // Доклады ТУСУРа. 2010. № 1(21). Ч. 1. С. 124–127.
11. *Басов О.О., Никитин В.В., Илюшин М.В.* Теоретико-множественная модель полимодальной инфокоммуникационной системы // Новые информационные технологии в научных исследованиях: материалы XVIII Всероссийской научно-технической конференции студентов, молодых ученых и специалистов. Рязанский государственный технический университет. 2013. С. 60–62.
12. *Басов О.О.* Математическая модель системы кодирования речевого сигнала с многопараметрической адаптацией // Телекоммуникации. 2008. № 7. С. 7–13.

References

1. Basov O.O., Saitov I.A. [Functioning quality and effectiveness of polymodal infocommunicational systems]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2014. vol. 1(32). pp. 152–170. (In Russ.).
2. Konceptual'nye polozhenija po postroeniju mul'tiservisnyh setej na VSS RF [Conceptual regulations concerning BCC RF multiservice networks]. M.: Minsvjazi RF. 2002. (In Russ.).
3. Ronzhin A.L., Karpov A.A., Li I.V. *Rechevoj i mnogomodal'nye interfejsy* [Speech and multimodal interfaces]. M.: Nauka, 2006. 173 p. (In Russ.).
4. Sheluhin O.I., Luk'jancev N.F. *Cifrovaja obrabotka i peredacha rechi: pod red. O. I. Sheluhina* [Digital processing and speech transmission. Edited by Sheluhin O.I.]. M.: Radio i svjaz'. 2000. 456 p. (In Russ.).
5. Kipjatkova I.S., Ronzhin A.L., Karpov A.A. *Avtomaticheskaja obrabotka razgovornoj russkoj rechi: monografija* [Automatic processing of Russian spoken voice: monographie]. SPb.: GUAP. 2013. 314 p. (In Russ.).
6. Gonsales R., Vuds R. *Cifrovaja obrabotka izobrazhenij* [Digital image processing]. M.: Tehnosfera, 2006. 1072 p. (In Russ.).
7. Ustinov A.A. *Stohasticheskoe kodirovanie video- i rechevoj informacii: monografija: v 2 ch. Pod red. prof. V.F. Komarovicha* [Video and spoken information stochastic coding: monographie in 2 parts. Edited by V.F. Komarovich]. SPb.: VAS. 2005. (In Russ.).
8. Stepanov S.N. *Osnovy teletrafika mul'tiservisnyh setej* [Teletraffic foundations of the multiservice networks]. M.: Jeko-Trendz. 2010. 392 p. (In Russ.).
9. Saitov I.A. *Osnovy teorii postroenija zashhishhennyh mul'tiprotokol'nyh opticheskih transportnyh setej telekommunikacionnyh system: monografija* [Theoretical bases of the secured multiprotocol optical transport networks construction of the telecommunicational systems monographie]. Orel: Akademija FSO Rossii. 2008. 220 p. (In Russ.).

10. Ronzhin A.L., Karpov A.A. [Design of interactive application with multimodal interface]. *Doklady TUSURa – Proceedings of TUSUR University*. 2010. vol. 1(21). part 1. pp. 124–127. (In Russ.).
11. *Basov O.O., Nikitin V.V., Iljushin M.V.* [Set-theoretical model of the polymodal infocommunicational system]. *Novye informacionnye tehnologii v nauchnyh issledovanijah: materialy XVIII Vserossijskoj nauchno-tehnicheskoj konferencii studentov, molodyh uchenyh i specialistov* [XVIII All-Russian scientific-technical conference of students, young scientists and specialists «New information technology in the scientific researches»]. Rjazanskij gosudarstvennyj tehnicheskij universitet. 2013. pp. 60–62. (In Russ.).
12. Basov O.O. [Mathematic model of the speech coding system with a multiparameter adaptation]. *Telekommunikacii – Telecommunications*. 2008. vol. 7. pp. 7–13. (In Russ.).

Басов Олег Олегович — к-т техн. наук, докторант, Академия Федеральной службы охраны Российской Федерации. Область научных интересов: обработка и кодирование речевых и иконических сигналов, проектирование полимодальных инфокоммуникационных систем. Число научных публикаций — 165. oobasov@mail.ru; Приборостроительная, 35, г. Орел, 302034; р.т.: 89192011897.

Basov Oleg Olegovich — Ph.D., doctoral student, The Academy of Federal Security Guard Service of the Russian Federation. Research interests: processing and coding of speech and iconic signals, polymodal infocommunicational systems design. The number of publications — 165. oobasov@mail.ru; 35, Priborostroitelnaya Street, Orel, 302034, Russia; office phone: 89192011897.

РЕФЕРАТ

Басов О.О. **Принципы построения полимодальных инфокоммуникационных систем на основе многомодальных архитектур абонентских терминалов.**

Решение прямой и двойственной задач синтеза ПИКС требует разработки теоретических конструкций (принципов построения, системы моделей) для оценки качества передачи полимодальной информации при заданных ресурсах СПД и их объема, требуемого для передачи максимального числа сообщений различных модальностей с заданным качеством.

В результате исследования установлено, что построение ПИКС должно базироваться на следующих основополагающих принципах: соответствия, системности, функциональной полноты, мультисервисности, полимодельности, открытости архитектуры, функциональной замкнутости, оперативной управляемости, дифференцируемости услуг и имеющихся ресурсов, децентрализации предоставляемых услуг (инвариантности доступа), ассоциативности и толерантности обращения к информации, гарантированного доступа к контексту, многооператорности и обратной связи.

Реализация перечисленных принципов позволяет строить инфокоммуникационные системы вне рамок традиционных принципов разделения передаваемой при межличностной коммуникации информации на услуги. Для количественного обоснования соответствующих решений была разработана специальная иерархическая система моделей ПИКС, ее подсистем и элементов, включающая в себя описательные, обобщенную и математические (макромодель, модели информационно-алгоритмической структуры, модели естественных и искусственных сигналов и услуг, модели кодирования и передачи полимодальной информации) модели, а также блок теоретических конструкций.

На основе разработанной системы моделей задача синтеза ПИКС сведена к поиску оптимальных (квазиоптимальных) подходов к формированию множества допустимых системотехнических решений с учетом ограничений на способы человеко-машинного взаимодействия и предоставляемые услуги, реализующих ПИКС на базе множества вариантов многомодальных интерфейсов абонентских терминалов.

На основе разрабатываемой теории ПИКС необходим обоснованный выбор конкретных вариантов, определяющих структуру и функции ПИКС и конфигурацию программно-аппаратного обеспечения, необходимого для ее реализации.

SUMMARY

Basov O.O. **Principles of Construction of Polymodal Info-Communication Systems based on Multimodal Architectures of Subscriber's Terminals.**

The polymodal infocomunicational system primal and dual synthesis task solution requires developing theoretical constructions (construction principles, model system) for polymodal information transmission quality assessment with the resources master data of the DTN and their volume necessary for transmitting the maximum number of the diverse modality messages with a stated quality.

In the course of the research it has been discovered that polymodal infocomunicational system must be based on the following core principles: accordance, consistency, functional completeness, multiservice, polymodality, architecture openness, functional closureness, operational manageability, services and given resources differentiation (access invariation), information inversion associativity and tolerance, guaranteed access to the context, multi-statement and inverse association.

Fulfilling of the stated principles allows building infocommunicational systems beyond the frames of the traditional principles of the interpersonal information transmission services division. For quantitative proof of the following solutions a special hierarchical model system of the polymodal infocommunicaional system, its subsystems and elements including descriptive, generalized and mathematical (macromodel, information-algorithmic structure model structure, natural and artificial signals and services models, coding and polymodal information transmission models) models and a theoretical construction block was developed.

According to the developed system of models the polymodal infocommunicational system synthesis problem is reduced to the optimal (quasioptimal) approaches of the acceptable circuit solutions multitude formation with the account of the limitation for the man-computer interaction and offered services which fulfill polymodal infocommunicational system based on the customer stations polymodal interfaces multitude variants.

Based on the developing polymodal infocommunicational systems theory a justified selection of the concrete variants defining system's structure and functions and also hardware configuration vital for its realization is required.

УДК 004.912

С.Н. Карпович

РУССКОЯЗЫЧНЫЙ КОРПУС ТЕКСТОВ SCTM-RU ДЛЯ ПОСТРОЕНИЯ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ

Карпович С.Н. **Русскоязычный корпус текстов SCTM-ru для построения тематических моделей.**

Аннотация. В статье рассматривается задача создания русскоязычного специального корпуса текстов для тестирования алгоритмов вероятностного тематического моделирования. В качестве наполнения корпуса предлагается использовать статьи международного новостного сайта «Русские Викиновости», распространяемого по свободной лицензии CC BY 2.5. Описан этап предварительной обработки и разметки корпуса текстов. Предложена разметка корпуса текстов, содержащая только необходимую в алгоритмах тематического моделирования информацию.

Ключевые слова: корпус текстов, обработка текста на естественном языке, тематическое моделирование, русский язык.

Karpovich S.N. **The Russian language text corpus for testing algorithms of topic model.**

Abstract. This paper describes the process of creating Russian language text corpus which is specialized for testing algorithms of probabilistic topic model. The articles of Wikinews licensed by Creative Commons Attribution 2.5 Generic (CC BY 2.5) were used as a source of texts for corpus. The stage of text's preprocessing and markup are described in the conclusion. We proposed an original markup of text corpus for testing algorithms of topic modeling.

Keywords: text corpora, topic model, natural language processing, Russian language.

1. Введение. В современном обществе главным продуктом и основным товаром становится информация. Активно развиваются: наука, экономика, политика, производственная сфера, во многих отраслях происходит создание и накопление цифровых данных. Для успешного извлечения и обработки информации из данных необходимо располагать подходящими инструментами, системами и алгоритмами. Растет потребность в системах обработки текстов на естественном языке. Обработка естественного языка (Natural Language Processing) уже применяется в привычных для пользователя программах и сервисах. Например, программы для чтения новостных лент умеют группировать новости по темам, поисковые системы находят документы с ценной для пользователя информацией, службы почтовых сообщений автоматически фильтруют спам. Используются различные алгоритмы кластеризации и классификации текстовых данных, наиболее популярные k-средних, SVM, нейронные сети. Перспективным направлением автоматической обработки текстов является разработка алгоритмов вероятностного тематического моделирования.

Тематическое моделирование – это способ построения тематической модели коллекции текстовых документов. Тематическая

модель задает отношение между темами и документами в корпусе текстов. Первое описание тематического моделирования появилось в работе Рагавана, Пападимитриу, Томаки и Вемполы 1998 году [1]. Томас Хофманн в 1999 году предложил вероятностное скрытое семантическое индексирование (PLSI) [2]. Одна из самых распространенных тематических моделей — это латентное размещение.

Дирихле (LDA), эта модель является обобщением вероятностного семантического индексирования и разработана Дэвидом Блейем, Эндрю Ыном и Майклом Джорданом в 2002 году [3]. Другие тематические модели, как правило, являются расширением LDA. В качестве примера на рисунке 1 представлен процесс построения тематической модели документа.

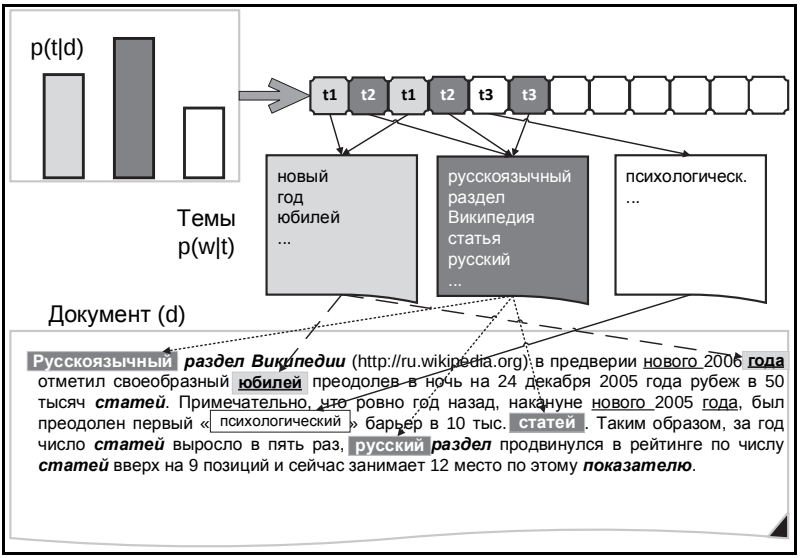


Рис. 1. Построение тематической модели документа: p(w|t) – матрица искомых условных распределений слов по темам, p(t|d) – матрица искомых условных распределений тем по документам: d — документ: w — слово: d, w — наблюдаемые переменные: t — тема (скрытая переменная)

Алгоритмы тематического моделирования ориентированы на работу с текстом на естественном языке. Первоначальные решения основывались на предположении, что текст — это «мешок слов», т.е. порядок слов в тексте не имеет значения. В последующих моделях успешно реализованы алгоритмы, учитывающие зависимости между словами с помощью скрытых Марковских моделей. В обзоре [4]

рассмотрены пять основных классов вероятностных тематических моделей: базовые, учитывающие отношения между документами, учитывающие отношения между словами, временные, обучаемые с учителем.

Наличие подготовленных текстовых корпусов позволит разрабатывать системы для автоматической обработки текстов на естественном языке, в том числе алгоритмы тематического моделирования. Создавая тематическую модель, необходимо учитывать языковые особенности текстов. Для развития методов тематического моделирования, работающих с русским языком, необходим русскоязычный корпус текстов, распространяемый по свободной лицензии.

Корпусная лингвистика – сложная лингвистическая дисциплина, которая сформировалась в последние десятилетия на базе электронной вычислительной техники. Она изучает построение лингвистических корпусов, способы обработки данных в них и собственно технологию их создания и использования. «Корпус – это информационно справочная система, основанная на собрании текстов на некотором языке в электронной форме», - такое определение текстового корпуса дано на сайте Национального Корпуса Русского Языка [5]. В научной работе [6] отмечено: «Корпусы, как правило, предназначены для неоднократного применения многими пользователями, поэтому их разметка и их лингвистическое обеспечение должны быть определенным образом унифицированы». Целесообразность создания и смысл использования корпуса определяется следующими предпосылками:

1. достаточно большой (репрезентативный) объем корпуса;
2. данные разного типа находятся в корпусе в своей естественной контекстной форме;
3. однажды созданный и подготовленный массив данных может использоваться многократно.

Корпусами первого порядка называют собрание текстов, объединенных общим признаком, например, источник, автор, место публикации. Специальный корпус текстов – это сбалансированный корпус, репрезентативный, как правило, небольшой по размеру подчиненный определенной исследовательской задаче и предназначенный для использования преимущественно в целях, соответствующих замыслу составителя. Текстовый корпус (text corpora, corpus), большая коллекция документов (large collection of documents), набор данных (dataset), как отмечено в работе [4] являются синонимичными понятиями.

Цель данной работы заключается в том, чтобы создать специальный русскоязычный корпус текстов СКТМ-ру (SCTM-ru), пригодный для исследования алгоритмов вероятностного тематического моделирования. Сформулируем требования к создаваемому корпусу. Корпус должен распространяться по свободной лицензии, количество документов должно быть достаточным для исследования, он должен содержать:

- оригинальный текст документов на естественном языке;
- даты описанных событий;
- информацию об авторстве;
- темы или тематические категории.

Рассмотрим возможность использования существующих текстовых корпусов для целей тестирования алгоритмов тематического моделирования.

2. Обзор текстовых корпусов и наборов данных. Национальный Корпус Русского Языка (НКРЯ) [5] – содержит более 335 тыс. документов на русском языке, разделенных на подкорпуса. Включает в себя 180 тыс. текстов Газетного корпуса. Для использования офлайновой версии основного корпуса (1 млн. словоупотреблений) необходимо подписать лицензионное соглашение. Статистика корпуса представлена на рисунке 2.

Подкорпус	Число текстов	Число предложений	Число словоупотреблений	% словоупотреблений
Основной корпус	76 882	17 574 752	209 198 275	57.3%
- в том числе со снятой омонимией	2 147	516 852	5 944 188	1.6%
Газетный корпус	181 175	8 553 495	113 292 003	31.0%
Диалектный корпус	197	20 273	194 283	0.1%
Обучающий корпус	229	65 666	664 751	0.2%
Параллельный корпус	370	1 609 609	24 022 437	6.6%
Поэтический корпус	41 448	638 861	6 738 474	1.8%
Устный корпус	3 034	1 604 626	10 122 579	2.8%
Мультимедийный корпус	31 741	148 619	648 576	0.2%
Всего:	**335 076**	**30 215 901**	**364 881 378**	**100%**

Рис. 2. Статистика НКРЯ, распределение текстов по подкорпусам [5]

Открытый корпус (OpenCorpora) [9, 10] – содержит порядка 3 тыс. документов на русском языке, 93 тыс. размеченных предложений,

10 источников данных, часть документов имеет информацию об авторе и дате описываемых событий. Частям текста приписана лингвистическая информация, такая как морфологическая, семантическая и синтаксическая. Для задач построения временных и автор-тематических моделей, корпус не подходит, так как не все документы корпуса содержат информацию об авторе и о дате событий.

Associated Press [11] – корпус содержит 2 тыс. документов на английском языке. Документы корпуса не имеют отметки о дате описанного события, авторе публикации, категории документа. Корпус применим для исследования ограниченного количества алгоритмов тематического моделирования.

The New York Times Annotated Corpus [12] – большой англоязычный текстовый корпус газетных заметок и новостей, распространяемый по закрытой лицензии.

20 Newsgroups [13] – коллекция новостей на английском языке, подготовленная для исследования алгоритмов автоматической обработки текстов. 20 новостных групп содержит порядка 20 тыс. документов. Важные для построения тематических моделей данные об авторе и дате публикации не размечены. Требуется предварительная обработка текста новостей для использования в тематическом моделировании.

Reuters Corpora [14] – большой англоязычный новостной корпус. В трех наборах данных более 3 млн. новостей. Распространяется по ограниченной лицензии только для научных исследований, предоставляется после подписания лицензионного соглашения. Существует более ранняя популярная для тестирования алгоритмов автоматической обработки текстов версия корпуса под названием Reuters-21578 [15], распространяемая по ограниченной лицензии, доступная для офлайн анализа.

Компьютерный корпус текстов русских газет конца XX-ого века [16, 17] – этот корпус был создан в 1999 году развивается и исследуется в настоящее время по грантам РФФИ. Корпус предназначен для анализа лингвистических особенностей (лексика, морфемика, морфология, словообразование, синтаксис, фразеология, стилистика) современного газетного языка. В корпусе 23 тыс. текстов по полным номерам 13-ти разных российских газет на русском языке. Размер корпуса 11 млн. словоупотреблений.

Корпус русского литературного языка [18, 19] – представлен в виде массива морфологически аннотированных текстов на русском литературном языке. Размер корпуса более 1 млн словоупотреблений со сбалансированным жанровым составом.

Хельсинкский аннотированный корпус русских текстов ХАНКО [20] – корпус содержит морфологическую, синтаксическую и функциональную информацию о текстах общим объемом 100 тыс. текстов, извлеченных из журнала «Итоги». Права на полные тексты статей журнала принадлежат правообладателям.

В таблице 1 представлено сравнение важных для исследования алгоритмов тематического моделирования характеристик корпусов: язык корпуса, лицензия распространения, доступность для скачивания и исследования на компьютерах без доступа в Интернет, информация об авторе, информация о дате описанных событий, тема текста.

Таблица 1. Сравнительная таблица характеристик текстовых корпусов

Корпус	Язык	Открытая лицензия	Доступен для скачивания	Инф. об авторе	Инф. о дате	Темы
НКРЯ	рус.	-	-	+	-	-
Открытый корпус	рус.	+	+	+	+	-
Associated Press	англ.	+	+	-	-	-
The New York Times Annotated Corpus	англ.	-	-	+	+	+
20 Newsgroups	англ.	+	+	-	-	-
Reuters Corpora	англ.	-	-	+	+	+
Компьютерный корпус текстов русских газет конца XX-ого века	рус.	-	-	-	-	-
Корпус русского литературного языка	рус.	-	-	-	-	-
ХАНКО	рус.	-	-	-	-	-
SCTM-ru	рус.	+	+	+	+	+

Рассмотренные текстовые корпуса в полной мере не соответствуют обозначенным в данной работе требованиям. Создаваемый специальный корпус для тематического моделирования СКТМ-ру (SCTM-ru) распространяется по свободной лицензии, язык корпуса русский, содержит информацию об авторстве, дате событий, тематической принадлежности документов, доступен для скачивания и проведения исследований на компьютерах без доступа в Интернет.

3. Технология создания корпуса SCTM-ru. Технологический процесс создания корпуса состоит из следующих шагов.

1. определение источника;
2. предварительная обработка текстов документов;
3. разметка параметров каждого документа в корпусе;
4. обеспечение доступа к корпусу.

В соответствии с обозначенным требованием к доступности данных корпуса, текста используемые в качестве наполнения должны

распространяться по свободной лицензии, должны быть доступны для скачивания и свободного использования.

В результате предварительной обработки текстов и разметки параметров каждого документа, в корпусе должна быть сохранена и специальным образом размечена информация необходимая для построения тематических моделей. Невостребованная в тематическом моделировании информация должна быть исключена из корпуса, за ненадобностью.

Различные задачи тематического моделирования могут требовать определенный порядок поступления данных в систему тематического моделирования, от последовательного для временных, до единовременного для обычных тематических моделей. Поэтому для обеспечения доступа к корпусу достаточно предоставить возможность для его скачивания, и последующего использования в соответствии с конкретными задачами, стоящими перед исследователем.

4. Источник данных для корпуса SCTM-ru. В качестве источника данных предлагаем использовать международный новостной сайт «Русские Викиновости» (Викиновости), тексты статей которого распространяются по свободной лицензии Creative Commons Attribution 2.5 Generic, доступны для скачивания и анализа на любых компьютерах, в том числе на компьютерах без доступа в Интернет. В работах [7, 8] отмечены преимущества вики-ресурсов, таких как Викисловарь и Википедия, для использования в качестве источника данных в исследовательских целях. Вики-ресурсы – это сайты второго поколения Интернет, характеризуемые тем, что к их наполнению привлечено огромное количество рядовых пользователей, с помощью которых происходит пополнение и актуализация информации. Большой объем, постоянное пополнение, нейтральность во взглядах, доступность относятся к преимуществам всех вики-ресурсов, в том числе к Викиновостям.

Викиновости – это братский проект большой Википедии, предназначен для написания новостных статей. Пример статьи Викиновостей представлен на рисунке 3. Отличительной особенностью сайта Викиновостей от любого другого новостного сайта является то, что каждый человек может принять участие в создании новости. Правила Викиновостей требуют писать новости с нейтральной точки зрения, в непредвзятом виде, выбирать существенные и актуальные темы, использовать достоверные источники.

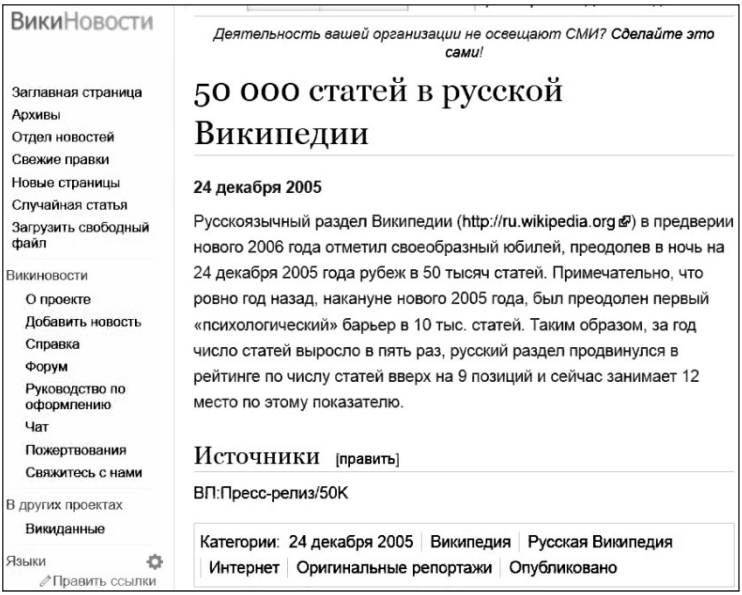


Рис. 3. Статья "50 000 статей в русской Википедии" на сайте русских Викиновостей

XML-файл экспорта базы данных Викиновостей состоит из следующих XML-элементов:

+ `<page>` – группа элементов новостной статьи;
+ `<title>` – название статьи;
– `<ns>` – идентификатор или имя пространства имен (namespace), элемент предназначен для отделения основных статей от служебных, ноль соответствует основному пространству имен;
+ `<id>` – уникальный идентификатор статьи;
+ `<revision>` – ревизия – это группа элементов актуальной версии статьи;
  – `<id>` – первичный ключ ревизии, используется для контроля изменений статьи;
  – `<parented>` – идентификатор родительской статьи;
  – `<timestamp>` – дата и время создания ревизии статьи;
  + `<contributor>` – группа элементов авторства статьи;
  – `<username>` – имя автора статьи;
  + `<id>` – уникальный идентификатор автора статьи;
  + `<text>` – текст статьи с элементами вики-разметки;

– `<sha1>` – хеш код статьи полученный алгоритмом криптографического хеширования SHA-1, используется для контроля версий;
– `<model>` – модель контента статьи, в данном случае wikitext;
– `<format>` – формат данных статьи, в данном случае text/x-wiki.

Для задач тематического моделирования необходима информация, которая содержится в элементах, отмеченных знаком плюс (+). Элементы, которые содержат информацию, неиспользуемую в алгоритмах тематического моделирования, отмечены знаком минус (-). Пример части XML-дерева экспортного файла базы данных Викиновостей представлен на рисунке 4.

```
<?xml version="1.0" encoding="utf-8"?>
<page>
  <title>50 000 статей в русской Википедии</title>
  <ns>0</ns>
  <id>1838</id>
  <revision>
    <id>75780</id>
    <parentid>39949</parentid>
    <timestamp>2011-10-01T18:45:01Z</timestamp>
    <contributor>
      <username>Schekinov Alexey Victorovich</username>
      <id>2156</id>
    </contributor>
    <text xml:space="preserve">{{:Дата |24 декабря 2005}}
{{ВикипедияН
 |Язык = Русская
 }}
Русскоязычный раздел [[Википедия|Википедии]] (http://ru.wikipedia.org)
в предверии нового 2006 года отметил своеобразный юбилей, преодолев в
ночь на 24 декабря 2005 года рубеж в 50 тысяч статей. Примечательно,
что ровно год назад, накануне нового 2005 года, был преодолен первый
«психологический» барьер в 10 тыс. статей. Таким образом, за год число
статей выросло в пять раз, русский раздел продвинулся в рейтинге по
числу статей вверх на 9 позиций и сейчас занимает 12 место по этому
показателю.
{{оригинальный репортаж 2}}
== Источники ==
{{w|ВП:Пресс-релиз/50K}}
{{публиковать}}
[[Категория:Русская Википедия]]</text>
    <sha1>ezckutzzznn6tioszytixr2atkv0v4p</sha1>
    <model>wikitext</model>
    <format>text/x-wiki</format>
  </revision>
</page>
```

Рис. 4. Пример XML статьи "50 000 статей в русской Википедии" на сайте русских Викиновостей

5. Предварительная обработка данных Викиновостей. В экспортном файле Викиновостей статьи отсортированы по дате создания ревизии `<timestamp>`, эта дата не связана с датой описанных событий. Авторам рекомендуется указывать с помощью вики-разметки дату событий в тексте статьи. Пример вики-разметки даты {{:Дата | 24 декабря 2005}} представлен на рисунке 4 внутри элемента `<text>`. Часть статей в экспортном файле Викиновостей не содержит дату событий в вики-разметке, но при этом она указана в тексте или в категории. Чтобы сохранить максимум востребованной в алгоритмах тематического моделирования информации, дата событий была по возможности восстановлена из текста и категорий. В 455 статьях не удалось восстановить дату событий, эти статьи являются подборками новостей, произошедших в один день, в разные годы и не представляют ценности для построения тематических моделей, они были исключены из корпуса. Документы корпуса SCTM-ru отсортированы по дате событий, от старых к новым.

В экспортном файле базы данных Викиновостей содержится информация об авторе последней ревизии статьи. Используем эту информацию как идентификатор авторства для построения автор-тематических моделей. Так как 58 Викиновостей не содержат информацию об авторе, а статьи ценны, то было принято техническое решение присвоить этим статьям уникальный идентификатор автора – 2, и включить их в корпус SCTM-ru.

Текст статьи Викиновостей содержит оформленные специальным образом ссылки. Ссылки делятся на три группы: внутренние – инструмент связывания страниц внутри языкового раздела Википедии, межязыковые ссылки (интервики) – средство для организации связей между различными вики-системами в сети Интернет и ссылки на страницы братских вики-проектов (например, на Википедию). Текст статьи, заключенный в двойные квадратные скобки является внутренней ссылкой, пример [[Википедия|Википедии]] представлен на рисунке 4. Если падеж ссылающегося слова или словосочетания не совпадает с именительным падежом, то в двойных квадратных скобках стоит черта, слева от которой указан именительный падеж текста ссылки, а справа текст, соответствующий грамматике предложения. Алгоритмы тематического моделирования учитывают количество вхождений каждой леммы слова в текст, во внутренних ссылках каждое слово имеет два вхождения в разных словоформах и будет дважды учтено в тематической модели, тем самым исказив частотные характеристики модели. В документах

корпуса SCTM-ru оставлена только та часть ссылки, которая соответствует грамматике предложения.

Новости должны сопровождаться ссылками на документальный источник. Они обычно делятся на четыре вида: другие статьи Викиновостей, внешние ссылки на онлайн-источники, цитаты печатных изданий и веб-сайты со справочной или связанной информацией. Для раздела статьи «Источники» используют вики-разметку == *Источники* == (см. пример на рисунке 4). Для целей тематического моделирования ссылки на источники не представляют большой ценности, поэтому было принято решение об их исключении из корпуса SCTM-ru.

Важным элементом разметки Викиновостей и важными данными для построения тематических моделей является информация о категориях, к которым статья имеет отношение. Категории статьи определяет ее автор.

Для предварительной обработки текстов была разработана программа на языке C#, среда разработки Visual Studio Express 2013. Для поиска по экспортному файлу Викиновостей использовались регулярные выражения. Пример задействованных регулярных выражений представлен в таблице 2. Программа многомодульная, каждый модуль выполняет одну определенную операцию. Программа получает на вход исходный XML-файл, специально подготовленные регулярные выражения последовательно обходят файл в поисках совпадения по шаблону, на выходе создается XML-файл с внесенными за одну итерацию изменениями. Для сохранения целостности первоначальных данных, каждый проход по исходному XML-файлу вносит лишь часть изменений, которые внимательно проверяет администратор системы, после чего программу запускают с другим модулем обработки.

Таблица 2. Примеры регулярных выражений для предварительной обработки текста Викиновостей

Регулярное выражение	Назначение поиска
^(=)?=(\s+)?Источник(и)?(\s+)?(=)?=\n(^([^\n]+)\n)+^\n	блок источники
\{\{Категории|([^|]+)|([^|]+)\}\}	блок категории
\[\[([^|]+)|([^\]]+)\]\]	Ссылки

Для подсчета статистики корпуса SCTM-ru была разработана многомодульная программа. Модуль подсчета документов осуществляет разбор XML-дерева корпуса, извлекает уникальные

идентификаторы каждого документа и считает их общее количество. Модуль подсчета авторов извлекает список уникальных идентификаторов авторов статей Викиновостей и подсчитывает их количество. Модуль подсчета категорий извлекает уникальные категории из XML-дерева корпуса и считает их количество. Модуль обработки дат описанных в статьях событий осуществляет разбор XML-дерева корпуса, извлекает информацию о дате события каждого документа, подсчитывает уникальные значения, находит самую раннюю и самую позднюю дату документа.

Для подсчета словарного состава корпуса SCTM-ru был разработан модуль с использованием регулярных выражений и программы MyStem. Модуль берет текст из заданных элементов XML-дерева (title, text), регулярные выражения из текста извлекают все последовательности букв русского алфавита. При подсчете слов последовательность букв русского алфавита, отделенная от других букв не буквами (например, знаки препинания, пробел), считается словом. Для определения лемм слов использовалась программа MyStem. Программа MyStem производит морфологический анализ текста на русском языке. Для слов, отсутствующих в словаре, порождаются гипотезы [21].

6. Разметка корпуса SCTM-ru. В качестве формата хранения корпуса SCTM-ru выбран XML (eXtensible Markup Language — расширяемый язык разметки), как один из наиболее удобных форматов для использования в программной среде и конвертации данных в другие форматы. Возможности XML позволяют сохранить текст исходной статьи Викиновости и выделить дополнительные параметры документа.

XML-файл корпуса (SCTM-ru) состоит из следующих элементов:

- `<page>` - группа элементов документа;
- `<title>` - название документа;
- `<id>` - уникальный идентификатор документа;
- `<userid>` - уникальный идентификатор автора;
- `<category>` - категория документа;
- `<date>` - дата событий документа;
- `<text>` - текст документа;

Пример разметки одного документа в корпусе SCTM-ru представлен на рисунке 5.

```
<?xml version="1.0" encoding="utf-8"?>
<page>
  <title>50 000 статей в русской Википедии</title>
  <id>1838</id>
  <userid>2156</userid>
  <category>Русская Википедия</category>
  <data>24 декабря 2005</data>
  <text>
    Русскоязычный раздел Википедии (http://ru.wikipedia.org) в предверии
    нового 2006 года отметил своеобразный юбилей, преодолев в ночь на 24
    декабря 2005 года рубеж в 50 тысяч статей. Примечательно, что ровно
    год назад, накануне нового 2005 года, был преодолен первый
    «психологический» барьер в 10 тыс. статей. Таким образом, за год число
    статей выросло в пять раз, русский раздел продвинулся в рейтинге по
    числу статей вверх на 9 позиций и сейчас занимает 12 место по этому
    показателю.
  </text>
</page>
```

Рис. 5. Пример XML-документа "50 000 статей в русской Википедии" в корпусе SCTM-ru

Заголовок документа (*title*) отделен от текста документа, т.к. словам заголовка может придаваться большее значение при построении тематической модели.

Уникальный идентификатор автора статьи (*userid*) – это параметр, который необходим в автор-тематических моделях. Автор-тематическая модель во времени (Author-Topic over Time) [22] представляет собой расширение LDA при построении модели оценивается распределение авторов, тем и документов по времени.

Категории документа (*category*) – это указанные автором статьи категории. Например, на рис. 4 в статье "50 000 статей в русской Википедии" указана категория «Русская Википедия». Информация о категории важна для тематического моделирования, поэтому сохранена в корпусе SCTM-ru см. рис. 5. Наличие информации о принадлежности документов к категориям позволит автоматически проверять точность, полноту, аккуратность тестируемых алгоритмов тематического моделирования. Информация о категориях документа может быть использована в моделях Labeled LDA, описанных в [23].

Дата описанных в статье событий (*date*) используется при построении временных (temporal) тематических моделей. Пример модели, использующей дату под названием «Тематики во времени» (Topic over Time - TOT) представлен в работе [24]. При построении временной модели наряду со стандартными распределениями слов по темам и тем по документам оцениваются

распределения каждой темы по времени, что позволяет проследить и отобразить динамику изменения тем во времени.

Текст документа (*text*) соответствует тексту исходной статьи. Мы целенаправленно оставляем исходный текст без изменения, без преобразования его в модель «мешка слов», без лингвистической обработки для возможности исследования уникальных особенностей русского языка. Информация о последовательности слов в тексте документа используется в моделях, учитывающих взаимную встречаемость слов. Например, модель под названием «Скрытая тематическая Марковская модель» (Hidden Topic Markov's Model - HTMM), описанная в работе [25], основана на предположениях, что слова в составе предложения, а также сами предложения связаны одной общей темой и темы слов в документе образуют цепь Маркова. В результате работы HTMM уменьшает неоднозначность слов, расширяет понимание темы.

7. Заключение. В результате проделанной работы был подготовлен специальный русскоязычный корпус текстов (SCTM-ru), подходящий для тестирования различных алгоритмов вероятностного тематического моделирования. Поставленные в работе цели были достигнуты: корпус SCTM-ru содержит оригиналы текстов документов на русском языке, информацию о дате описанных в документе событиях, информацию об авторе и категориях, к которым относится документ, доступен для скачивания и использования на устройствах без доступа в Интернет.

Источником данных корпуса является международный новостной сайт «Русские Викиновости». Корпус SCTM-ru состоит из 7 тыс. документов, 185 авторов, почти 12 тыс. уникальных категорий. События, описанные в документах, распределены по более чем 2 тыс. уникальным датам, с ноября 2005 года по июнь 2014 года. В корпусе SCTM-ru 2,4 млн словоупотреблений, состоящих только из русских букв. Словарный состав корпуса – 150,6 тыс. уникальных словоформ, 59 тыс. уникальных лемм.

Объем созданного корпуса дает основания предположить его репрезентативность для различных задач автоматической обработки текстов на естественном языке. Как отмечено в работе [26] «Неразумно ждать пока кто-то по-научному сбалансирует корпус, перед тем как его использовать, и неосмотрительно было бы оценивать результаты анализа корпуса как «малодостоверные» или «неуместные» просто потому, что нельзя доказать, что используемый корпус «сбалансирован»». Разнообразие описанных в корпусе SCTM-ru событий и огромный коллектив авторов статей (21 тыс. участников)

обосновывают предположение о его сбалансированности. Убедиться в сбалансированности корпуса можно после проведения анализа его внутренних признаков и построения тематических моделей.

Предложенная технология создания корпуса текстов для задач тематического моделирования позволяет расширять корпус SCTM-ru за счет новых статей. Аналогичным образом могут быть созданы языковые корпуса на любом из 33-х представленных в Викиновостях языках. В предложенном формате могут быть созданы коллекции и корпуса, из различных источников данных, при этом должна быть сохранена только востребованная в алгоритмах тематического моделирования информация.

Далее на базе созданного корпуса будут исследованы особенности существующих вариаций алгоритмов тематического моделирования, будут разработаны новые алгоритмы, учитывающие лингвистические особенности русского языка. Корпус SCTM-ru распространяется по открытой лицензии, доступен для скачивания на сайте www.cims.ru.

Литература

1. *Papadimitriou C.H., Raghavan P., Tamaki H., Vempala S.* Latent semantic indexing: A probabilistic analysis. 1998. pp. 159–168.
2. *Hoffman T.* Probabilistic Latent Semantic Indexing // Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval. 1999. pp. 50–57.
3. *Blei D.M., Ng A.Y., Jordan M.I.* Latent Dirichlet Allocation // Journal of Machine Learning Research. 2003. pp. 993–1022.
4. *Daud A., Li J., Zhou L., Muhammad F.* Knowledge discovery through directed probabilistic topic models: a survey // In Proceedings of Frontiers of Computer Science in China. 2010. pp. 280–301.
5. Сайт Национального корпуса русского языка НКРЯ. URL: www.ruscorpora.ru. (дата обращения: 12.01.2015).
6. *Захаров В.П.* Международные стандарты в области корпусной лингвистики // Структурная и прикладная лингвистика. 2012. № 9. С. 201–221.
7. *Крижановский А.А., Смирнов А.В.* Подход к автоматизированному построению общецелевой лексической онтологии на основе данных викисловаря // Известия РАН. Теория и системы управления. 2013. № 2. С. 53–63.
8. *Смирнов А.В., Круглов В.М., Крижановский А.А., Луговая Н.Б., Карпов А.А., Кипяткова И.С.* Количественный анализ лексики русского WordNet и викисловарей // Труды СПИИРАН. 2012. Вып. 23. С. 231–253.
9. *Грановский Д.В., Бочаров В.В., Бичинева С.В.* Открытый корпус: принципы работы и перспективы // Компьютерная лингвистика и развитие семантического поиска в Интернете: Труды научного семинара XIII Всероссийской объединенной конференции «Интернет и современное общество». Санкт-Петербург. 2010 г. СПб. 2010. 94 с.
10. Сайт Открытого корпуса. URL: opencorpora.org (дата обращения: 10.01.2015).
11. Small corpus of Associated Press. URL: www.cs.princeton.edu/~blei/lda-c/ (дата обращения: 06.01.2015).

12. The New York Times Annotated Corpus. URL: catalog.ldc.upenn.edu/LDC2008T19 (дата обращения: 14.01.2015).
13. The 20 Newsgroups data set. URL: qwone.com/~jason/20Newsgroups/ (дата обращения: 24.01.2015).
14. Reuters Corpora. URL: trec.nist.gov/data/reuters/reuters.html (дата обращения: 24.01.2015).
15. Reuters-21578 Text Categorization Collection Data Set. URL: archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection (дата обращения: 24.01.2015).
16. *Виноградова В.Б., Кукушкина О.В., Поликарпов А.А., Савчук С.О.* Компьютерный корпус текстов русских газет конца 20-го века: создание, категоризация, автоматизированный анализ языковых особенностей // "Русский язык: исторические судьбы и современность" Международный конгресс русистов-исследователей. Труды и материалы. М.: Изд-во Моск. ун-та. 2001. С. 114–115.
17. Компьютерный корпус текстов русских газет конца XX-ого века. URL: www.philol.msu.ru/~lex/corpus/corp_descr.html (дата обращения: 24.01.2015).
18. *Венцов А.В., Грудева Е.В.* О корпусе русского литературного языка (narusco.ru) // Русская Лингвистика. 2009. Том 33. № 2. С. 195–209.
19. Корпус русского литературного языка. URL: www.narusco.ru (дата обращения: 24.01.2015).
20. Хельсинкский аннотированный корпус русских текстов ХАНКО. URL: www.helsinki.fi/venaja/russian/e-material/hanco/index.htm (дата обращения: 24.01.2015).
21. Официальный сайт программы морфологического анализа текстов на русском языке MyStem. URL: api.yandex.ru/mystem/ (дата обращения: 12.12.2014).
22. *Xu S., Shi Q., Qiao X., et al.* Author-Topic over Time (AToT): a dynamic users' interest model, in Mobile, Ubiquitous, and Intelligent Computing // Springer. Berlin. 2014. pp. 239–245.
23. *Ramage D., Hall D., Nallapati R., Manning C.D.* Labeled LDA. A supervised topic model for credit attribution in multi-labeled corpora // In Empirical Methods in Natural Language Processing. 2009. pp. 248–256.
24. *Wang X., McCallum A.* Topics over Time: A Non-Markov Continuous Time Model of Topical Trends // In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia. USA. 2006.
25. *Gruber A., Rosen-Zvi M., Weiss Y.* Hidden Topic Markov Models. In: Proceedings of Artificial Intelligence and Statistics (AISTATS) // San Juan. Puerto Rico. USA. 2007.
26. *Захаров В.П., Азарова И.В.* Параметризация специальных корпусов текстов // Структурная и прикладная лингвистика: Межвузовский сборник. СПб: СПбГУ. 2012. Вып. 9. С. 176–184.

References

1. Papadimitriou C.H., Raghavan P., Tamaki H., Vempala S. Latent semantic indexing: A probabilistic analysis. 1998. pp. 159–168.
2. Hoffman T. Probabilistic Latent Semantic Indexing. Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval. 1999. pp. 50–57.
3. Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003. pp. 993–1022.
4. Daud A., Li J., Zhou L., Muhammad F. Knowledge discovery through directed probabilistic topic models: a survey. In Proceedings of Frontiers of Computer Science in China. 2010. pp. 280–301.

5. Sajt Nacional'nogo korpusa russkogo jazyka NKRJa. [Website of Russian National Corpus]. Available at: www.ruscorpora.ru (accessed: 12.01.2015). (In Russ.).
6. Zakharov V.P. [International standards in corpora linguistics]. *Strukturnaja i prikladnaja lingvistika – Structural and Applied Linguistics*. 2012. vol. 9. pp. 201–221. (In Russ.).
7. Smirnov A.V., Krizhanovsky A.A. [An approach to automated construction of a general-purpose lexical ontology based on Wiktionary]. *Izvestija RAN. Teorija i sistemy upravlenija – Journal of Computer and Systems Sciences International.* 2013. vol. 52. no. 2. pp. 215–225. (In Russ.).
8. Smirnov A.V., Kruglov V. M., Krizhanovsky A.A., Lugovaya N.B., Karpov A.A., Kipyatkova I.S. [A quantitative analysis of the lexicon in Russian WordNet and Wiktionaries]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2012. vol. 23. pp. 231–253. (In Russ.).
9. Granovsky D.V., Bocharov V.V., Bichineva S.V. [Opencorpora: how it work and perspectives]. *Kompyuternaya lingvistika i razvitie semanticheskogo poiska v internete: Trudy nauchnogo seminara XIII vserossiyskoy obedinennoy konferencii «internet i sovremennoe obschestvo»* [Computer linguistics and development of semantic search on Internet: Proceedings of the 13th All-Russian integrated conference «Internet and Modern Society»]. St. Petersburg. 2010. 94 p. (In Russ.).
10. Sajt Otkrytogo korpusa [OpenCorpora Website]. Available at: opencorpora.org (accessed: 15.01.2015). (In Russ.).
11. Small corpus of Associated Press. Available at: www.cs.princeton.edu/~blei/lda-c/ (accessed: 6.01.2015).
12. The New York Times Annotated Corpus. Available at: catalog.ldc.upenn.edu/LDC2008T19 (accessed: 14.01.2015).
13. The 20 Newsgroups data set. Available at: qwone.com/~jason/20Newsgroups/ (accessed: 24.01.2015).
14. Reuters Corpora. Available at: trec.nist.gov/data/reuters/reuters.html (accessed: 24.01.2015).
15. Reuters-21578 Text Categorization Collection Data Set. Available at: archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection (accessed: 24.01.2015).
16. Vinogradova V.B., Kukushkina O.V., Polikarpov A.A., Savchuk S.O. [The computer corpus of Russian newspapers of the XX th century end: the creation, categorization, automated analysis of linguistic features]. *"Russkiy yazyk: istoricheskie sudby I sovremennost." Mezhdunarodnyy congress rusistov issledovateley. Moskva, filologicheskiy f t MGU im. M.V. Lomonosova* ["Russian Language: its Historical Destiny and Present State" International Congress of Russian Language Researchers]. M.: University Pressю 2001. pp. 114–115. (In Russ.).
17. Komp'juternyj korpus tekstov russkih gazet konca XX-ogo veka [The computer corpus of Russian newspapers of the XXth century end]. Available at: www.philol.msu.ru/~lex/corpus/corp_descr.html (accessed: 24.01.2015). (In Russ.).
18. Vencov A.V., Grudeva E.V. [About Corpus of Standard Written Russian (narusco.ru)]. *Russkaja Lingvistika – Russian Linguistics.* 2009. vol. 33. no. 2. pp. 195–209. (In Russ.).
19. Korpus russkogo literaturnogo jazyka [Corpus of Standard Written Russian]. Available at: www.narusco.ru (accessed: 24.01.2015). (In Russ.).
20. Hel'sinkskij annotirovannyj korpus russkih tekstov HANKO [HANCO Corpus]. Available at: www.helsinki.fi/venaja/russian/e-material/hanco/index.htm (accessed: 24.01.2015).

21. Oficial'nyj sajt programmy morfologicheskogo analiza tekstov na russkom jazyke MyStem [System for automatic morphological analysis of Russian MyStem]. Available at: api.yandex.ru/mystem/ (accessed: 12.12.2014). (In Russ).
22. Xu S., Shi Q., Qiao X., et al. Author-Topic over Time (AToT): a dynamic users' interest model, in Mobile, Ubiquitous, and Intelligent Computing. Springer. Berlin. 2014. pp. 239–245.
23. Ramage D., Hall D., Nallapati R., Manning C.D. Labeled LDA. A supervised topic model for credit attribution in multi-labeled corpora. In Empirical Methods in Natural Language Processing. 2009. pp. 248–256.
24. Wang X., McCallum A. Topics over Time: A Non-Markov Continuous Time Model of Topical Trends. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia. USA. 2006.
25. Gruber A., Rosen-Zvi M., Weiss Y. Hidden Topic Markov Models. In Proceedings of Artificial Intelligence and Statistics (AISTATS). San Juan. Puerto Rico. USA. 2007.
26. Zakharov V.P., Azarova I.V. [Special text corpora parametrization]. *Strukturnaya i prikladnaya lingvistika: mezhvuzovskiy sbornik* – Structural and Applied Linguistics: Interuniversity collection. SPb. St. Petersburg State University. 2012. vol. 9. pp 176–184. (In Russ.).

Карпович Сергей Николаевич — руководитель отдела поисковой оптимизации ООО "Рамблер Интернет Холдинг". Область научных интересов: тематическое моделирование, обработка текстов на естественном языке, кластеризация, классификация, обработка данных, машинное обучение. Число научных публикаций — 1. cims@yandex.ru, http://www.cims.ru/; 117105, Москва, Варшавское ш., 9, стр. 1, БЦ «Даниловская мануфактура», корпус «Ряды Солдатенкова»; р.т. +7(495)7851700.

Karpovich Sergey Nikolaevich — head of Search Engine Optimization Rambler Internet Holding LLC. Research interests: topic model, natural language processing, classification, clustering, data mining. The number of publications — 1. cims@yandex.ru, http://www.cims.ru/; Varshavskoe sh., 9, str. 1, BC «Danilovskaja manufaktura», korpus «Rjady Soldatenkova», 117105, Moscow; office phone +7(495)7851700.

РЕФЕРАТ

Карпович С.Н. **Русскоязычный корпус текстов SCTM-ru для построения тематических моделей**

В статье предложен специальный корпус текстов для тестирования алгоритмов тематического моделирования SCTM-ru. В условиях стремительного роста количества информационных данных, остро проявляется проблема разработки инструментов и систем для их автоматической обработки. Для создания систем и тестирования алгоритмов должны существовать подходящие наборы данных. Необходимо наличие свободных коллекций документов, текстовых корпусов на русском языке, для исследований методов автоматической обработки текстов на естественном языке, с учетом лингвистических особенностей языка. Обозначены требования к специальному корпусу: он должен распространяться по свободной лицензии, количество документов должно быть достаточным для исследования, должен содержать текста документов на естественном языке, должен содержать востребованную в алгоритмах тематического моделирования информацию. Проведен сравнительный анализ корпусов на русском и иностранных языках, выявлено несоответствие характеристик существующих корпусов с обозначенным требованиям.

Описана технология создания корпуса, выбор подходящего источника данных, этап предварительной обработки текстов документов, разметка корпуса и обеспечение доступа. Источником данных корпуса является международный новостной сайт «Русские Викиновости». Корпус SCTM-ru состоит из 7009 документов, 185 авторов, 11 895 уникальных категорий. События, описанные в документах, распределены по 2 236 уникальным датам, с ноября 2005 года по июнь 2014 года. В корпусе SCTM-ru 2,4 млн словоупотреблений, состоящих только из русских букв. Словарный состав корпуса – 150,6 тыс. уникальных словоформ, 59 тыс. уникальных лемм. Корпус репрезентативен. Убедиться в сбалансированности корпуса предлагается в ходе его исследования.

Разработанный подход создания корпуса позволяет постоянно расширять корпус SCTM-ru за счет новых статей. Аналогичным образом может быть подготовлен языковой корпус на любом из 33-х представленных в Викиновостях языках. Предложенная технология подготовки корпуса текстов для задач тематического моделирования позволяет создавать коллекции и корпуса из данных, полученных из различных источников, при этом будет сохранена только востребованная в алгоритмах тематического моделирования информация. Далее на базе созданного корпуса будут исследованы особенности существующих вариаций алгоритмов тематического моделирования, будут разработаны новые алгоритмы, учитывающие лингвистические особенности русского языка.

SUMMARY

Karpovich S.N. **The Russian language text corpus for testing algorithms of topic model.**

This paper proposes a special text corpus SCTM-ru to be used for testing algorithms of probabilistic topic model. With rapidly increasing amounts of information, there is a critical need for tools and systems to be able to automatically process them. To create the systems and to test the algorithms require suitable data sets. We need free document libraries, text corpora in Russian to research into the methods of natural language processing taking into account the linguistic features of the language. The requirements for a special corpus have been defined: it should be distributed under a free license, contain enough documents for the research, with the text in the documents being in natural languages, and contain relevant information for the topic modelling. A comparative analysis of corpora in Russian and foreign languages has been done revealing a non-compliance of the existing corpora with the above requirements.

The article describes a technology for the creation of such a corpus, how to choose a suitable data source, a stage for the preprocessing of the texts, and how to format and provide access to the corpus. The data source for the corpus is the international news website Russian Wikinews. The SCTN-ru corpus contains 7,009 documents written by 185 authors and split into 11,895 unique categories. The events described in the documents cover 2,236 unique dates, from November 2005 to June 2014. There are 2.4 million tokens consisting only in Russian letters in SCTM-ru corpus. The corpus contains 150.6 thousand unique word forms and 59 thousand unique lemmas. The corpus is representative. The readers are invited to see that the corpus is balanced during its analysis.

The developed approach to the creation of the corpus allows SCTM-ru to be constantly expanded with new articles. In a similar way, a corpus in any of 33 languages presented in Wikinews can be created. The proposed technology of the creation of text corpora for the topic modelling makes it possible to create collections and corpora using data obtained from different sources, with only relevant for the topic model information being saved. Next, the created corpus will be used as a basis for the research into the features of the existing variations of the topic modelling, and new algorithms taking into account the linguistic features of Russian language will be developed.

УДК 57.087

М.А. Бланк, О.А. Бланк, Е.М. Мясникова, С.Б. Рудницкий, Д.М. Денисова

ОСОБЕННОСТИ РАСПРЕДЕЛЕНИЯ ИНТЕГРАТИВНЫХ ПОКАЗАТЕЛЕЙ ТРЕВОЖНОСТИ ОНКОЛОГИЧЕСКИХ БОЛЬНЫХ, ВЫЯВЛЕННЫЕ СТАТИСТИЧЕСКИМИ СПОСОБАМИ

Бланк М.А., Бланк О.А., Мясникова Е.М., Рудницкий С.Б., Денисова Д.М. **Особенности распределения интегративных показателей тревожности онкологических больных, выявленные статистическими способами.**

Аннотация. В данной статье представлены результаты исследования, в ходе которого были выявлены ранее неизвестные особенности распределения показателей ситуативной тревожности у больных злокачественными новообразованиями, проходящих курс противоопухолевой терапии. На основе статистических методов показано, что в процессе постановки диагноза и лечения распределение меняет свой тип от бимодального к унимодальному, с возвращением к бимодальному распределению при достижении клинической ремиссии. В результате нами сделан важный вывод о несоответствии распределения показателей во всех исследуемых группах нормальному распределению, что накладывает определенные ограничения на применение статистических методов. Расчеты статистических критериев произведены с помощью пакета Statistica 6.0 и функции dip.test на языке R.

Ключевые слова: ситуативная тревожность, личностная тревожность, распределение показателей, злокачественные новообразования.

Blank M.A., Blank O.A., Myasnikova E.M., Rudnitsky S.B., Denisova D.M. **Distribution Patterns of Integrated Anxiety Rates in Cancer Patients Revealed by Statistical Tools.**

Abstract. This article presents research results that show previously unknown patterns of state anxiety rate distribution in patients with malignant neoplasms undergoing antitumor therapy. Statistical methods have shown that distribution type changes from bimodal to unimodal during diagnosis and treatment and returns back to a bimodal pattern once clinical remission is achieved. As a result, we have come to an important conclusion that rates in all examined samples are not normally distributed, which places certain limitations upon the use of statistical methods. Statistica 6.0 package and dip.test in the R language were used for performing statistical tests.

Keywords: state anxiety, trait anxiety, rate distribution, malignant neoplasms.

1. Введение. Личностная тревожность, присущая конкретному индивидууму, является одним из основных показателей, отражающих психологические особенности человека [1–3]. Уровень личностной тревожности на протяжении жизни меняется незначительно. На изменения внешних обстоятельств человек реагирует изменением уровня ситуативной тревожности (тревоги). Известно, что к повышению уровня ситуативной тревоги приводит появление какой-либо опасности, например, значимой угрозы жизни, возможности снижения или потери трудоспособности, нарушения социальной адаптации, ухудшения или потери материального благополучия, сопряженных с тяжелой

болезнью. В ряду таких заболеваний одно из лидирующих мест занимают злокачественные новообразования. Исследователи, изучавшие тревожность онкологических больных, утверждали, что для них характерен повышенный уровень ситуативной тревожности с момента постановки диагноза до окончания лечения [4–6], а после окончания лечения он снижается [7–9].

Целью подавляющего большинства исследований, предпринимаемых онкологами, является повышение эффективности лечения больных. Сохранение удовлетворительного качества жизни пациента в период проведения курса лечения злокачественного новообразования и по его завершении является такой же приоритетной целью исследований в области онкологии, как повышение эффективности противоопухолевых воздействий.

Критериями, определяющими эффективность лечения онкологических больных, традиционно служат непосредственные и отдаленные результаты проведенного хирургического вмешательства, лекарственной противоопухолевой терапии или лучевой терапии.

Непосредственные результаты лечения определяют, оценивая степень регресса опухоли, и констатируют регресс злокачественного новообразования, стабилизацию или прогрессирование процесса. Средняя продолжительность жизни пациентов, медиана выживаемости и выживаемость в течение определенного срока (процент больных, проживших от момента установления диагноза, например, 1, 3, 5 или 10 лет) являются основными параметрами, используемыми в оценке отдаленных результатов противоопухолевого воздействия.

Качество жизни – интегративный показатель, характеризующий все многообразные аспекты витальных проявлений человека и как биологического объекта, и как личности. Уровень этого показателя обеспечивают не только квалифицированные медицинские воздействия и правильная организация ухода за больным, но и мероприятия, направленные на трудовую, социальную, психологическую, бытовую и семейную реабилитацию пациента. Таким образом, качество жизни индивидуума зависит как от наличия и выраженности соматических страданий (болевого синдрома, нарушения функций органов и систем), от степени нарушения физической и умственной активности, обеспечивающих сохранение или утрату трудоспособности и возможности самообслуживания, от социальной поддержки и адаптации, так и от психологического статуса человека.

Известно, что психологическое состояние пациента не только в значительной мере определяет качество его жизни, но и влияет на эффективность проводимого лечения [10–12]. В перечне известных ныне

факторов, от которых может зависеть результат специфических противоопухолевых воздействий [13], отведено немаловажное место именно психологическим факторам. Таким образом, можно утверждать, что психологические особенности пациента могут влиять на эффекты проводимой терапии. С другой стороны, нельзя исключить, что степень эффективности лечения в конечном итоге влияет на психологическое состояние пациента.

Одним из трендов развития современной медицины стало повсеместное внедрение стандартов оказания медицинской помощи, четко регламентирующих объем и характер медицинских услуг. «Персонализация» лечебного воздействия в онкологии сводится к формированию и реализации плана лечения с учетом определенных характеристик опухоли (вплоть до молекулярно-генетических). В перечень параметров, определяющих выбор того или иного стандарта лечения, не входят индивидуальные особенности конкретного организма и совокупность сопутствующей патологии. Необходимость коррекции, соответствующей психотипу каждого пациента и его психологическим реакциям на заболевание и на связанные с ним медицинские вмешательства, также выпадает из поля зрения врача. Таким образом, «индивидуализация» и «персонализация» лечения в онкологии носят сугубо декларативный характер.

Рациональная организация психологического сопровождения пациентов возможна не только при условии правильного представления об изменениях психологического статуса больного на разных этапах течения опухолевого процесса, но и от типа психологической акцентуации пациента (в момент установления первичного диагноза, во время проведения всего курса лечения, в период ремиссии, в случаях рецидива заболевания, в терминальной стадии). Эти знания необходимы также для оптимального построения взаимоотношений «больной – врач» [10].

Статистические исследования в области психологии и математическая обработка полученных данных сопряжены с определенными трудностями, поскольку количественную оценку психологических особенностей личности и реакций на изменение ситуации производят с использованием так называемых «мягких» шкал. Тем более важным становится корректный выбор методов статистики для получения результатов, адекватно характеризующих изучаемые нематериальные параметры и позволяющих делать правильные выводы и заключения.

Целью исследования явилось изучение возможных особенностей психологического статуса онкологических больных, находящихся

в процессе лечения. Достижение поставленной цели обеспечивалось решением следующих задач:

1) определением показателей тревожности практически здоровых лиц;

2) изучением тревожности беременных женщин. Беременность – это нормальное особое состояние организма, во время которого происходит развитие эмбриона и впоследствии плода. У женщин, находящихся в этом состоянии, как и у ряда онкологических больных, повышается содержание в крови раково-эмбрионального антигена (РЭА), поэтому они были выбраны в качестве группы отрицательного контроля;

3) установлением показателей тревожности больных злокачественными новообразованиями, получающих специфическую противоопухолевую терапию;

4) определением показателей тревожности онкологических больных, введенных в длительную (≥4-5 лет) клиническую ремиссию.

2. Материалы и методы. Ранее в процессе выполнения плановой научной работы, посвященной созданию инструмента для оценки психосоматического состояния человека [14, 15], нами была создана обширная база данных, позволившая выполнить настоящее исследование.

Материалом для исследования послужили данные компьютерного опроса состояния тревожности в группах, соответствующих поставленным задачам, а именно:

1) практически здоровых взрослых людей, прошедших полноценное обследование во время диспансеризации (99 человек, 103 исследования);

2) беременных женщин (171 женщина, 196 исследований);

3) больных злокачественными новообразованиями различных локализаций (23 человека, 39 исследований), находящихся в процессе специфического лечения;

4) больных злокачественными новообразованиями (42 человека, 79 исследований), введенных в длительную клиническую ремиссию.

Все исследуемые были подвергнуты тестированию с использованием Интегративного теста тревожности (ИТТ), компьютерного варианта для взрослых, разработанного в Психоневрологическом институте им. В.М. Бехтерева [16, 17]. Тест отражает такие составляющие, как неспецифический эмоциональный дискомфорт, астенический компонент, фобический компонент, тревожную оценку перспективы и компонент социальной защиты. Выбор Интегративного теста тревожности обусловлен тем, что ИТТ позволяет количественно оценить пе-

речисленные проявления личностной тревожности и ситуативной тревоги. Тест понятен и прост, анкеты опросника доступны для самостоятельного заполнения испытуемыми, обладающими минимальной компьютерной грамотностью. Обработка данных производится автоматически. Все обследование занимает не более 10-15 минут и может быть реализовано на амбулаторном врачебном приеме в любом медицинском учреждении.

Две исследовательские и две контрольные группы были сформированы в соответствии с дизайном планового исследования. В первую исследовательскую группу вошли больные с морфологически верифицированным диагнозом злокачественного новообразования (злокачественные лимфомы и солидные опухоли), проходившие тестирование в период проведения специфического противоопухолевого лечения. Вторую исследовательскую группу составили пациенты, введенные в длительную (не менее 4 лет) клиническую ремиссию и находящиеся под динамическим наблюдением в Федеральном государственном бюджетном учреждении «Российский научный центр радиологии и хирургических технологий». Эти лица не реже одного раза в год подвергались полному физикальному, клинико-лабораторному и лучевому обследованию, подтверждавшему ремиссию заболевания. Подавляющая часть группы представлена больными злокачественными лимфомами. Группа положительного контроля сформирована из практически здоровых волонтеров, прошедших углубленную диспансеризацию и получивших допуск к работе с источниками ионизирующего излучения. В группу отрицательного контроля были включены практически здоровые беременные женщины, обследованные и находящиеся под наблюдением в одной из женских консультаций Санкт-Петербурга.

Расчеты статистических критериев произведены с помощью пакета Statistica 6.0 [18] и функции dip.test на языке R [19].

3. Специфика распределения показателей тревожности в разных группах. Сравнительный анализ гистограмм распределения интегративных показателей тревожности в четырех группах позволил сделать следующие выводы. В двух группах – здоровых лиц и беременных женщин – показатели как ситуативной тревоги, так и личностной тревожности демонстрируют распределение бимодального типа (рисунки 1.1, 1.2), которое, естественно, не является «нормальным». Похожую картину можно наблюдать и в группе больных злокачественными новообразованиями, введенных в длительную клиническую ремиссию (четвертая группа) (рисунок 1.4). Анализируя характер рас-

пределения обоих показателей интегративной тревожности в этой группе, также можно сделать вывод о его бимодальном характере.

Совершенно иную тенденцию демонстрируют распределения тех же показателей в группе больных, находящихся в процессе специфического противоопухолевого лечения по поводу злокачественных новообразований (третья группа). Если для показателей личностной тревожности, как и во всех предыдущих группах, наблюдается бимодальный характер распределения, то у распределения показателей ситуативной тревожности очевидно наличие одного ярко выраженного максимума (рисунок 1.3, а).

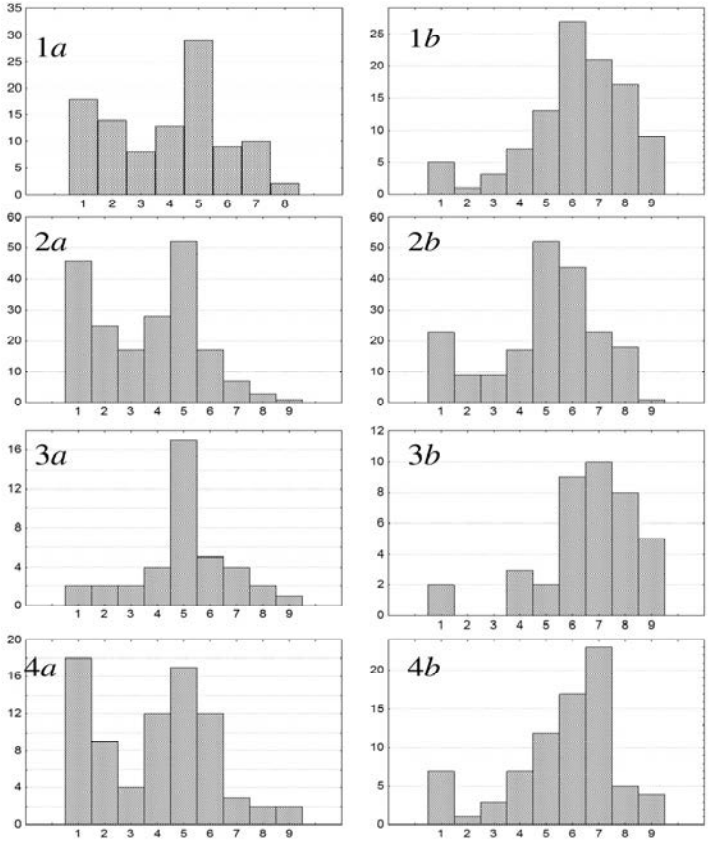


Рис.1. Гистограммы распределения интегративных показателей ситуативной (a) и личностной (b) тревожности у здоровых взрослых людей – 1, беременных женщин – 2, онкологических больных, проходящих лечение – 3 и онкологических больных, введенных в устойчивую клиническую ремиссию – 4

Выводы, сделанные на основании визуального анализа гистограмм показателей тревожности, подтверждаются с помощью статистических методов. Во-первых, мы провели попарное сравнение всех четырех групп с помощью критерия Колмогорова-Смирнова. Тест показывает значимое ($p<0.05$) отличие показателей ситуативной тревожности третьей группы от остальных групп, причем все остальные группы значимо не различаются между собой ($p>0.1$ для всех попарных сравнений). (Замечание: Если принять во внимание поправку Бонферрони на множественные сравнения (4 независимых сравнения), то отличие третьей группы от первой и четвертой значимо на уровне $p<0.1$.) Во-вторых, был проведен DIP-тест на унимодальность распределения [20], который показал, что гипотеза унимодальности отвергается с уровнем значимости $p<10^{-7}$ для всех групп, кроме третьей. Величина p-значения для третьей группы составляет 0.28, что позволяет сделать вывод об унимодальном распределении показателей.

Поскольку унимодальность является необходимым, но не достаточным условием нормальности распределения, показатели тревожности в третьей группе были дополнительно протестированы на нормальность распределения с помощью критерия Колмогорова-Смирнова. Таким образом, гипотеза нормальности была отвергнута с уровнем значимости $p<0.01$.

Отклонение от нормальности в данном случае наглядно характеризуется значением коэффициента эксцесса, который для третьей группы равен 0.66, что значительно больше нуля. Большие положительные значения коэффициента свидетельствуют об островершинности распределения по сравнению с нормальным, что и можно наблюдать для гистограммы третьей группы на рисунке 1.3, а.

4. Заключение. Во всех исследованных группах распределение интегративных показателей как ситуативной тревоги, так и личностной тревожности не соответствует закону нормального (Гауссова) распределения. Ввиду этого использование статистических методов, требующих нормальности распределения вектора данных, является некорректным и может приводить к формированию ошибочных представлений.

При этом нами показано, что во всех группах, кроме одной, распределение обладает свойством бимодальности. Только распределение интегративных показателей ситуативной тревоги в группе больных, получающих лечение, носит ярко выраженный унимодальный характер. Показатели ситуативной тревоги пациентов, которые после завершения противоопухолевой терапии были введены в длительную ремиссию, вновь распределяются бимодально. Выявленный нами фе-

номен является неожиданным, поскольку противоречит сложившемуся представлению о повышении тревожности у всех онкологических пациентов.

Можно предположить, что обнаруженное в результате нашего исследования бимодальное распределение интегративных показателей личностной тревожности и ситуативной тревоги у людей, пребывающих в различных жизненных ситуациях, таких как здоровье и активная трудовая деятельность, ожидание ребенка, восстановление активности и адаптации после успешного проведения курса противоопухолевого лечения при наличии диагноза онкологического заболевания, отражает распределение общеизвестных психотипов человека в популяции.

Феномен изменения показателей ситуативной тревоги в группе онкологических больных, пребывающих в процессе лечения, мы объясняем изменением реакции на реальную и значимую угрозу жизни вследствие разделения или перемещения ответственности за исход на других лиц (лечащего врача, другой медицинский персонал).

Изучение причин этого явления лежит в сфере интересов и профессиональной деятельности клинических психологов. Полагаем, что установленная нами закономерность «усреднения» показателей ситуативной тревоги распространяется не только на больных злокачественными новообразованиями, получающих квалифицированное лечение, но и на пациентов, страдающих от неонкологических заболеваний, угрожающих их жизни, и имеющих возможность разделить ответственность за исход болезни с медицинским персоналом.

Литература

1. *Zeidner M., Matthews G.* Anxiety 101 // Springer Publishing Company. 2010. 180 p.
2. *Eysenck M.* Anxiety and Cognition: A Unified Theory // Psychology Press. 2014. 209 p.
3. Anxiety: Current Trends in Theory and Research / Edited by Spielberger C.D. // Academic Press. 2013. 268 p.
4. *Козлова Н.В., Андросова Т.В.* Социально-психологическое сопровождение онкологических больных // Вестник Томского государственного университета. 2010. № 335. С. 142–147.
5. *Квасова Е.В., Новицкий А.В., Апчел В.Я., Гордиенко А.В., Павлова Н.В., Сухонос Ю.А., Дорохов Г.В., Пятибрат Е.Д.* Особенности психологического статуса и самооценки качества жизни у больных со злокачественными лимфомами при различной химиотерапевтической тактике // Вестник Российской военно-медицинской академии. 2013. № 1(41). С. 131–135.
6. *Jadoon N.A., Munir W., Shahzad M.A., Choudhry Z.S.* Assessment of depression and anxiety in adult cancer outpatients: a cross-sectional study // BMC Cancer. 2010. vol. 10. 594 p.
7. *Ольшевская Н.С., Гуменюк Л.Н., Прохоров Д.В.* Анализ психологических характеристик личности, тревожности и депрессии у пациентов с меланоцитарными новообразованиями кожи // Медічна психология. 2013. №2 С. 19–22.

8. *Семионкин Е.И., Яковлева Н.В., Трушин С.Н.* Исследование некоторых аспектов психического статуса больных колоректальным раком // Российский медико-биологический вестник. 2010. №3. С. 23–27.
9. *Bulotiene G., Veseliunas J., Ostapenko V., Furmonavicius T.* Women with breast cancer: relationships between social factors involving anxiety and depression // Archives of Psychiatry and Psychotherapy. 2008. no. 4. pp. 57–62.
10. *Гнездилов А.В.* Путь на Голгофу // СПб.: Фирма «КЛИНТ». 1995. 136 с.
11. *Колосов А.Е., Шиповников Н.Б.* Психологические нарушения при диагнозе «рак» // Киров: «Вятка». 1994. 136 с.
12. *Мишин Ю.Б.* Философия рака. Этиология, патогенез, лечение и профилактика: Заметки практикующего врача-онколога // Волгоград: Волгоградское научное издательство. 2005. 108 с.
13. *Бланк М.А., Бланк О.А.* Хронобиомедицина для онкологии // СПб.: NIKA. 2010. 120 с.
14. *Бланк М.А., Рудницкий С.Б., Бланк О.А., Вассерман Е.Л., Вильнер Г.А., Жвалевский О.В., Денисова Д.М., Ширяева О.А.* К оценке психосоматического статуса человека // Материалы научной конференции «От лучей Рентгена – к инновациям XXI века: 90 лет со дня основания первого в мире рентгенорадиологического института (Российского научного центра радиологии и хирургических технологий)» с участием специалистов стран ближнего и дальнего зарубежья. Санкт-Петербург. 2008. С. 343.
15. *Юсупов Р.М., Вассерман Е.Л., Денисова Д.М., Дюк В.А., Жвалевский О.В., Карташев Н.К., Рудницкий С.Б., Толстоногов Д.А., Бланк М.А., Бланк О.А.* Разработка теоретических основ, моделей и информационных технологий экспресс-диагностики и мониторинга функционального состояния человека на основе комплексной обработки биометрических данных // РАН Программа фундаментальных исследований Президиума РАН. Фундаментальные науки – медицине. М.: Фирма «Слово». 2009. С. 105–106.
16. *Бизюк А.П., Вассерман Л.И., Иовлев Б.В.* Применение интегративного теста тревожности (ИТТ). Методические рекомендации // СПб.: Изд-во НИПНИ им. В. М. Бехтерева. 2003. 23 с.
17. *Червинская К.Р., Щелкова О.Ю.* Медицинская психодиагностика и инженерия знаний / Под ред. Л.И.Вассермана // СПб.: Ювента; М.: Издательский центр «Академия». 2002. 624 с.
18. Официальный сайт компании StatSoft Russia. URL: http://www.statsoft.ru/ (дата обращения: 19.02.2015).
19. Hartigan's dip test statistic for unimodality. URL: http://cran.r-project.org/web/packages/diptest/index.html (дата обращения: 19.02.2015).
20. *Hartigan J.A., Hartigan P.M.* The DIP test of unimodality // The Annals of Statistics. 1985. vol. 13. pp. 70–84.

References

1. Zeidner M., Matthews G. Anxiety 101. Springer Publishing Company. 2010. 180 p.
2. Eysenck M. Anxiety anad Cognition: A Unified Theory. Psychology Press. 2014. 209 p.
3. Anxiety: Current Trends in Theory and Research. Edited by Spielberger C.D. Academic Press. 2013. 268 p.
4. Kozlova N.V., Androsova T.V. [Social and psychological support of cancer patients]. *Vestnik Tomskogo gosudarstvennogo universiteta – Bulletin of Tomsk state university*. 2010. no. 335. pp. 142–147. (In Russ.).

5. Kvasova E.V., Novitskiy A.V., Apchel V.Ya., Gordienko A.V., Pavlova N.V., Sukhonos Yu.A., Dorokhov G.V., Pyatibrat E.D. [Psychological status features and life quality self-evaluation in patients with malignant lymphoma treated with different chemotherapy approach]. *Vestnik Rossiyskoy voenno-meditsinskoy akademii – Bulletin of Russian Military Medical Academy.* 2013. no. 1(41). pp. 131–135. (In Russ.).
6. Jadoon N.A., Munir W., Shahzad M.A., Choudhry Z.S. Assessment of depression and anxiety in adult cancer outpatients: a cross-sectional study. BMC Cancer. 2010. vol. 10. 594 p.
7. Olshevskaya N.S., Gumeniuk L.N., Prokhorov D.V. [The analysis of psychological characteristics of personality, anxiety, and depression in patients with melanocytic neoplasms of the skin]. *Medichna psikhologiya – Medical psychology*. 2013. no. 2 pp. 19–22. (In Russ.).
8. Semionkin E.I., Yakovleva N.V., Trushin S.N. [Investigation of some aspect of psychic status in patients with colorectal cancer]. *Rossiyskiy mediko-biologicheskiy vestnik – Russian Medical Biological Herald.* 2010. no 3. pp. 23–27. (In Russ.).
9. Bulotiene G., Veseliunas J., Ostapenko V., Furmonavicius T. Women with breast cancer: relationships between social factors involving anxiety and depression. *Archives of Psychiatry and Psychotherapy.* 2008. no 4. pp. 57–62.
10. Gnezdilov A.V. *Put' na Golgofu* [Path to Calvary]. SPb.: Firma «KLINT». 1995. 136 p. (In Russ.).
11. Kolosov A.E., Shipovnikov N.B. *Psihologicheskiye narushenia pri diagnoze "rak"* [Psychological disorders and the diagnosis of cancer]. Kirov: «Viatka». 1994. 136 p. (In Russ.).
12. Mishin Yu.B. *Filosofiya raka. Etiologiya, patogenez, lechenie I profilaktika. Zametki praktikuyuschego vracha-onkologa* [The philosophy of cancer. Aetiology, pathogenesis, treatment and prophylaxis. Notes of a practicing oncologist]. Volgograd: Volgogradskoye nauchnoye izdatel'stvo. 2005. 108 p. (In Russ.).
13. Blank M.A., Blank O.A. *Khronobiomeditsina dlya onkologii* [Chronobiomedicine for oncology]. Saint Petersburg: NIKA. 2010. 120 p. (In Russ.).
14. Blank M.A., Rudnitsky S.B., Blank O.A., Wasserman E.L., Vilner G.A., Zhvalevsky O.V., Denisova D.M., Shiryaeva O.A. [On evaluating human psychosomatic state]. *Materialy nauchnoy konferentsii "Ot luchey Rentgena – k innovatsiyam XXI veka: 90 let so dnya osnovaniya pervogo v mire rentgenoradiologicheskogo instituta (Rossiyskogo nauchnogo tsentra radiologii i khirurgicheskikh tekhnologiy)» s uchastiem spetsialistov stran blizhnego I dal'nego zarubezh'ya* [From Roentgen rays to XXI century innovations: 90 years since the foundation of the world's first roentgenological and radiological institute (Russian Research Centre for Radiology and Surgical Technologies): Collected papers]. Saint Petersburg. 2008. pp. 343. (In Russ.).
15. Yusupov R.M., Wasserman E.L., Denisova D.M., Duke V.A., Zhvalevsky O.V., Kartashev N.K., Rudnitsky S.B., Tolstonogov D.A., Blank M.A., Blank O.A. [Development of theoretical foundations, models and information technologies for express diagnosis and monitoring of human functional state on the basis of integrated biometrical data processing]. *RAN. Programma fundamental'nykh issledovaniy Prezidiuma RAN. Fundamental'nye nauki – meditsine* [RAS Presidium's program of fundamentsl research. Fundamental science for medicine]. Moscow: Firma "Slovo". 2009. pp. 105–106. (In Russ.).
16. Bizyuk A.P., Wasserman L.I., Iovlev B.V. *Primeneniye integrativnogo testa trevozhnosti (ITT). Metodicheskiye rekomendatsii* [Use of integrated anxiety test (IAT). Manual]. SPb.: Izd-vo NIPNI im. V. M. Bekhtereva, 2003. 23 p. (In Russ.).
17. Chervinskaya K.R., Schelkova O.Yu. *Meditsinskaya psihodiagnostika i inzheneriya znaniy: pod red. L.I.Vassermana* [Medical psychodiagnostics and knowledge engi-

neering. Edited by L.I.Wasserman]. SPb.: Yuventa; M.: Izdatelski tsentr “Akademiya”. 2002. 624 p. (In Russ.).
18. Oficial'nyj sajt kompanii StatSoft Russia [Official web site of StatSoft Russia company]. Available at: http://www.statsoft.ru/ (accessed 19.02.2015). (In Russ.).
19. Hartigan's dip test statistic for unimodality. Available at: http://cran.r-project.org/web/packages/diptest/index.html (accessed 19.02.2015).
20. Hartigan J.A., Hartigan P.M. The DIP test of unimodality. *The Annals of Statistics.* 1985, vol. 13, pp. 70–84.

Бланк Михаил Аркадьевич — д-р мед. наук, руководитель научной группы "Хрономедицина", Федеральное государственное бюджетное учреждение «Российский научный центр радиологии и хирургических технологий» Министерства здравоохранения Российской Федерации (ФГБУ РНЦРХТ). Область научных интересов: онкология, хрономедицина, радиология. Число научных публикаций — 149. mablank@mail.ru; ул. Ленинградская д. 70, п.Песочный, Санкт-Петербург, 197758, РФ; р.т.: +7 (812) 234-54-14.

Blank Mikhail Arkadievich — Dr. Sci., head of Chronomedicine research group, Federal government-financed research establishment, Russian Research Centre for Radiology and Surgical Technologies (RRCRST). Research interests: oncology, chronomedicine, radiology. The number of publications — 149. mablank@mail.ru; 70 Leningradskaya st., Pesochny, Saint Petersburg, 197758, Russia; office phone: +7 (812) 234-54-14.

Бланк Ольга Алексеевна — д-р мед. наук, ведущий научный сотрудник научной группы "Хрономедицина" отдела клинической радиологии, Федеральное государственное бюджетное учреждение «Российский научный центр радиологии и хирургических технологий» Министерства здравоохранения Российской Федерации (ФГБУ РНЦРХТ). Область научных интересов: онкология, радиология, хрономедицина. Число научных публикаций — 65. oablank@mail.ru; ул. Ленинградская д. 70, п.Песочный, Санкт-Петербург, 197758, РФ; р.т.: 89119461511.

Blank Olga Alexeevna — Dr. Sci., leading researcher of Chronomedicine research group, Federal state-financed establishment, Russian Research Centre for Radiology and Surgical Technologies (RRCRST). Research interests: oncology, radiology, chronomedicine. The number of publications — 65. oablank@mail.ru; 70 Leningradskaya st., Pesochny, Saint Petersburg, 197758, Russia; office phone: 89119461511.

Мясникова Екатерина Марковна — к-т техн. наук, ведущий научный сотрудник, Санкт-Петербургский государственный политехнический университет (СПбПУ). Область научных интересов: математическая статистика, биоинформатика, системная биология. Число научных публикаций — 40. ekmyasnikova@yandex.ru; Политехническая 29, Санкт-Петербург, 195251; р.т.: +79219574375.

Myasnikova Ekaterina Markovna — Ph.D., leading researcher, Saint-Petersburg State Polytechnical University. Research interests: statistics, bioinformatics, systems biology. The number of publications — 40. ekmyasnikova@yandex.ru; 29, Politekhnicheskaya str., St.Petersburg, 195251, Russia; office phone: +79219574375.

Рудницкий Сергей Борисович — д-р техн. наук, заведующий лабораторией биомедицинской информатики, Федеральное государственное бюджетное учреждение науки Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН). Область научных интересов: комплексная обработка сигналов, принятие решений в условиях неопределенности, биометрия, дальняя радионавигация. Чис-

ло научных публикаций — 95. sbr@spiiras.ru; 14 линия В.О., 39, Санкт-Петербург, 199178, РФ; р.т.: +7(812) 328-54-11, Факс: +7(812) 328-44-50.

Rudnitsky Sergey Borisovich — Ph.D., Dr. Sci., head of the laboratory of biomedical informatics, Federal government-financed research establishment, St. Petersburg institute of informatics and automation of Russian Academy of Sciences (SPIIRAS). Research interests: integrated signal processing; decision making under conditions of uncertainty; biometry; long-range radionavigation. The number of publications — 95. sbr@spiiras.ru; 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone: +7(812) 328-54-11, Fax: +7(812) 328-44-50

Денисова Дарья Михайловна — младший научный сотрудник лаборатории биомедицинской информатики, Федеральное государственное бюджетное учреждение науки Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН). Область научных интересов: исследование эмоциональной сферы человека, разработка психологических методов моделирования эмоционально значимых ситуаций, психофизиология стресса, поведение, ориентированное на выживание. Число научных публикаций — 11. dendm@spiiras.ru; 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ; р.т.: +7(812)328-54-11.

Denisova Daria Mikhailovna — junior researcher of the laboratory of biometrical informatics, Federal government-financed research establishment, St. Petersburg institute of informatics and automation of Russian Academy of Sciences (SPIIRAS). Research interests: investigations in human emotional sphere, design of psychological modelling methods of emotion-inducing situations, psychophysiology of stress, survival-oriented behaviour. The number of publications — 11. dendm@spiiras.ru; 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone: +7(812)328-54-11.

Поддержка исследований. Работа выполнена при финансовой поддержке РФФИ (проект № 15-04-07800-а); ОНИТ РАН (программа ОИ6).

Acknowledgements. This research is supported by RFBR (grant 15-04-07800-a); DNIT RAS (program OI6).

РЕФЕРАТ

Бланк М.А., Бланк О.А., Мясникова Е.М., Рудницкий С.Б., Денисова Д.М. **Особенности распределения интегративных показателей тревожности онкологических больных, выявленные статистическими способами.**

Эффективность лечения онкологических больных в определенной степени зависит от такого фактора, как психологическое состояние пациента. Одним из основных показателей, отражающих психологическое состояние человека, является тревожность.

Ранее, при выполнении плановой научной работы, посвященной созданию инструмента для оценки психосоматического состояния человека, нами была создана обширная база данных, позволившая выполнить настоящее исследование, целью которого явилось изучение возможных особенностей психологического статуса онкологических больных, находящихся в процессе лечения.

Были исследованы показатели тревожности следующих групп: практически здоровых лиц, беременных женщин, больных злокачественными новообразованиями, находящихся в процессе специфического лечения, и больных злокачественными новообразованиями, введенных в длительную клиническую ремиссию.

В результате исследования было выявлено следующее: показатели личностной тревожности во всех группах имеют тенденцию к бимодальному распределению. То же касается и показателей ситуативной тревожности у всех групп, кроме группы больных, находящихся в процессе лечения. У них в отношении характера распределения показателей ситуативной тревожности очевидна ярко выраженная тенденция к формированию унимодального распределения.

Таким образом, во всех исследованных группах распределение показателей как ситуативной, так и личностной тревожности, проявляя бимодальность, не соответствует закону нормального (Гауссова) распределения, что свидетельствует о некорректности использования методов, требующих нормальности распределения вектора данных. Как следствие, применение неадекватных статистических подходов может приводить к формированию ошибочных представлений.

SUMMARY

Blank M.A., Blank O.A., Myasnikova E.M., Rudnitsky S.B., Denisova D.M.
Distribution Patterns of Integrated Anxiety Rates in Cancer Patients Revealed by Statistical Tools.

To a certain extent the effectiveness of treatment of cancer patients depends on factors such as the patient's psychological state. One of the principal rates describing human psychological state is the anxiety rate.

Earlier, as we carried out planned research to create an instrument for evaluating human psychosomatic state, we had created a vast database that had allowed us to perform this study. Its goal was investigating into possible patterns of psychological state in cancer patients undergoing treatment.

We have studied the anxiety rates in the following groups: people with no evidence of disease, pregnant women, patients with malignant neoplasms undergoing specific treatment, and patients with malignant neoplasms brought into lasting clinical remission.

Results of our study have shown the following: trait anxiety rates in all the groups have a tendency for bimodal distribution. The same applies to state anxiety rates in all the groups, the sole exception being the group of patients undergoing treatment. In their case the distribution pattern of state anxiety rates shows a pronounced tendency for the forming of unimodal distribution.

Thus, in all the studied groups the rate distribution for both state and trait anxiety shows bimodality, which means that it does not conform to the law of normal (Gaussian) distribution. Therefore, it's incorrect to apply methods that demand the distribution normality of the data vector. The use of inadequate statistical approaches can therefore result in the forming of erroneous concepts.

УДК 511; 519.6

А.Е. Ваулин, М.С. Назаров

СВЕДЕНИЕ ЗАДАЧИ ФАКТОРИЗАЦИИ НАТУРАЛЬНОГО ЧИСЛА К ЗАДАЧЕ РАЗБИЕНИЯ ЧИСЛА НА ЧАСТИ. ЧАСТЬ 1

Ваулин А.Е., Назаров М.С. **Сведение задачи факторизации натурального числа к задаче разбиения числа на части. Часть 1.**

Аннотация. В настоящей работе рассматриваются вопросы разработки алгоритмов факторизации составных натуральных чисел. Анализ возможностей существующих алгоритмов показывает, что в перспективе ближайших десятилетий существенного прогресса в повышении их быстродействия ожидать не приходится. Дело, по-видимому, в ограниченности одностороннего математического подхода, базирующегося на использовании математических решет. Автором предлагается иной подход, основанный на изучении внутренней структуры натурального ряда чисел и использовании свойств чисел, не зависящих от их разрядности (по типу признаков делимости).

Ключевые слова: натуральный ряд, нечетное число, ф-инвариант числа, разбиения числа, контур натурального ряда чисел.

Vaulin A.E., Nazarov M.C. **Conversion of Integer Factorization to a Problem of Decomposition of a Number. Part 1.**

Abstract. The development of factorization mechanisms of composite integer numbers is considered in this work. The existent methods will not become more rapid and efficient in the nearest decade, due to narrow and inadequate mathematical approach to solution of this problem, which is based on so-called sieve of Eratosthenes. The mechanism suggested by author of this work, uses a completely new method based on examination of internal structure of natural sequence and application of digit place independent features (the criterion for divisibility).

Keywords: natural number, odd number, f-invariant of a numbers, partitions of a number, time-beating, natural numbers circuit.

1. Введение. В работе анализируются возможности факторизации больших натуральных чисел, и показывается необходимость разработки новых методов решения этой задачи за приемлемые для практических нужд временные интервалы. В общей постановке проблема факторизации является проблемой теории чисел, так как среди арифметических операций этой теории отсутствует операция факторизации натурального числа [8–12], которая удовлетворяла бы запросам науки и общественной практики.

Обращается внимание на исключительно важную роль нечетных натуральных чисел на их свойства и особенности. Показывается, что алгоритмы факторизации сегодняшнего уровня развития теории самым тесным образом связаны со свойствами чисел, зависящими от разрядности факторизуемых чисел. Такая зависимость не позволяет создать быстродействующие алгоритмы факторизации [2–7]. В работе предлагается опираться на свойства чисел свободные от такой зависи-

мости, и разработать алгоритмы свободные от нее. О существовании таких свойств свидетельствуют известные признаки делимости.

В работе намечается путь ослабления и даже полного устранения связи, определяющей длительность выполнения факторизации с разрядностью числа. С этой целью используются новые установленные свойства нечетных натуральных чисел (ННЧ), не зависящие от их разрядности. Длительность вычислений при этом требуется существенно меньшая и слабо зависит от разрядности числа.

Приводятся описания двух моделей ННЧ: *интервальной* и *нумерационной*, а также показывается, как используются понятия модели натурального ряда чисел.

На основе нумерационной модели и теоремы о предельном контуре [4] разработан алгоритм факторизации произвольных ННЧ, использующий формирование разбиения (представление суммой номеров контуров) половины номера $k_п(N)/2$ предельного контура числа N. Поскольку значение $k_п(N)/2$ – ограниченное число, и возможности его представления суммой подряд следующих чисел конечны, то отсюда следует конечность алгоритма и его сходимость. Числовые примеры иллюстрируют основные понятия моделей, их взаимосвязи и демонстрируют работоспособность предлагаемых методов.

Приводится возможный алгоритм факторизации числа, использующий связь кубов чисел и сумм ННЧ, легко вычисляемых в рамках модели натурального ряда чисел (НРЧ).

Формула, описывающая интервал длиной N, расстоянием между граничными точками интервала (между квадратами), реализует разложение числа N на множители, т.е. реализует его факторизацию.

Изложение материала сопровождается многочисленными ссылками на публикации и числовыми примерами, призванными сделать его более ясным и доступным.

2. Проблема факторизации больших чисел. В теории чисел (высшей арифметике) настоящего времени отсутствует простая и доступная операция (факторизация) обратная умножению чисел – разложение составного числа на множители. Отдельные числа большой, но ограниченной разрядности с большими трудностями удается разложить квалифицированным специалистам, но в принципе задача сегодня из разряда нерешаемых.

Задача факторизации известна с древнейших времен, как задача разложения натурального числа на простые множители, но до настоящего времени она не получила практически полезного результативного разрешения. Самыми известными результатами на сегодняшний день в области создания метода решения задачи факторизации боль-

ших чисел (ЗФБЧ) следует признать методы и алгоритмы различных математических решет. Теория решет берет свое начало от решета Эратосфена (до н.э.), позднее придуманы решета Бруна, Сельберга, Линника, а последнее достижение – это решето с числовым полем, предложенное в 1990 году Х.В. Ленстра, А.К. Ленстра, Манассе и Поллардом.

В лучших традициях 17 века, когда отдельные математики (Ферма, Мерсенн и др.) формулировали математические задачи и в личной переписке предлагали их для решения коллегам за рубежом и в своей стране, поступила фирма RSA.

Фирма в 1991 году представила на своем сайте в интернете список из 42 чисел [10], которые предложила факторизовать любому желающему, испытать свои силы и возможности на этом поприще. Достижения человечества в решении задачи факторизации хорошо иллюстрируются данными таблицы 1.

Таблица 1. Достижения в области факторизации больших чисел, список которых объявлен фирмой RSA в 1991году

Число	Количество десятичных цифр	Стоимость	Дата факторизации
RSA – 100	100		Апрель 1991
RSA – 110	110		Апрель 1992
RSA – 120	120		Июнь 1993
RSA – 129	129	$100	Апрель1994
RSA – 130	130		Апрель 10, 1996
RSA – 140	140		Февраль 2, 1999
RSA – 150	150		Апрель 16, 2004
RSA – 155	155		Август 22, 1999
RSA – 160	160		Апрель 1, 2003
RSA – 200	200		Май 9, 2005
RSA – 576	174	$10 000	Декабрь 3, 2003
RSA – 640	193	$20 000	Ноябрь 4, 2005
RSA – 704	212	$30 000	–
RSA – 768	232	$50 000	Январь 2010
RSA – 896	270	$75 000	–
RSA – 1024	309	$100 000	–
RSA – 1536	463	$150 000	–
RSA – 2048	617	$200 000	–

Для подкрепления интереса к поиску решений заданий из списка фирма назначила премии за правильно найденное решение для отдельных чисел. Таблица 1 содержит 18 чисел из этого списка RSA. Часть чисел этого списка уже факторизована, но с момента опубликования прошло уже более 20 лет.

Видим, что за 20 с небольшим лет лучшими математиками преодолен рубеж факторизации для конкретного числа только из 232 десятичных цифр. Другое число такой же разрядности потребует для факторизации не намного меньшее время. Заметим также, что каждое из чисел списка формировалось как произведение всего лишь двух простых чисел практически одинаковой разрядности. Эта дополнительная информация, возможно, способствует поиску решения.

По-видимому, алгоритмы, используемые математиками для факторизации, существенным образом зависят от разрядности факторизуемого числа. Такой вывод следует из рассмотрения таблицы. За меньшее время (исчисляемое в годах) были разложены числа меньшей разрядности. Использование мультипликативной модели числа приводит к огромному перебору вариантов, хотя такой перебор, конечно же, не является тотальным. Размер области поиска решения с течением времени (в годах) очень медленно сокращается.

В предлагаемой работе рассматривается другой, оригинальный подход к решению задачи факторизации, который опирается на модели натурального ряда чисел в целом и отдельного натурального числа.

Исключительно важную роль в рассматриваемом подходе играют некоторые теоремы, натуральные нечетные числа и, в частности, последовательности нечетных чисел, для которых вводятся классы.

Теорема (*Основная теорема арифметики*):

Каждое целое число, неравное нулю, представляется произведением степеней простых чисел единственным образом с точностью до порядка сомножителей и их знаков

$$n = \prod_i p_i^{\alpha_i} = p_1^{\alpha_1} p_2^{\alpha_2} \ldots p_k^{\alpha_k} \cdot$$

Эта формула представляет каноническое разложение числа n на сомножители.

В основной теореме арифметики выделяют два утверждения, требующие доказательства. Во-первых, утверждение о существовании представления всякого целого числа произведением степеней простых чисел, и во-вторых, утверждение о единственности такого представления. Доказательства обоих утверждений приводятся практически во всех руководствах и учебниках по теории чисел. Результат этой теоремы достигается для произвольного числа процедурой факторизации.

Теорема (Факторизация натуральных чисел): Произвольное составное натуральное число N может быть представлено произведением чисел (факторизовано) путём последовательного выполнения над ним следующих преобразований:

1. Если N – составное чётное натуральное число, то оно представляется в виде $N = 2^{t_2} \cdot p_2$,

где $p_2 \equiv 1 (\mathrm{mod}\ 2)$ – нечётное число, $t_2 = 1(1)\ldots$, и $2 \nmid p_2$;

2. Если $N = p_2$ – нечётное число, оканчивающееся цифрой **5**, то оно представляется в виде $N = 5^{t_5} \cdot p_5$, где p_5 – нечётное число, $t_5 = 1(1)\ldots$; и $5 \nmid p_5$;

3. Если $N = p_5$ - нечётное число, оканчивающееся одной из цифр **1, 3, 7, 9**, а его свёртка $s(N)$ (сумма цифр) кратна числу **3**, то оно представляется в виде $N = 3^{t_3} \cdot p_3$,

где p_3 – нечётное число, $t_3 = 1(1)\ldots$; и $3 \nmid p_3$;

4. Если $N = p_3$ - нечётное число, оканчивающееся одной из цифр **1, 3, 7, 9**, то оно имеет вид $N = p_k + 30 \cdot t$, где t – натуральное число, а $p_k \in \{7, 11, 13, 17, 19, 23, 29, 31\}$, и факторизацию можно выполнить, например, с использованием теоремы о предельном контуре.

Поскольку делители составного нечетного натурального числа (СННЧ) – это натуральные числа, то, по-видимому, можно предположить, что сами делители некоторого натурального числа $N = d_м\ d_м = p_1^{a_1} \cdot p_2^{a_2} \cdot \ldots \cdot p_k^{a_k}$ и их кратные значения некоторым образом распределены в натуральном ряде чисел. Эти числа (делители и кратные им) содержат информацию о делителях N. Очевидно, что кратных значений может быть бесконечно много, но практически в работе будут использоваться только меньшие N значения. Будем далее рассматривать задачу факторизации СННЧ N.

Возможны несколько вариантов информированности исследователя задачи относительно делителей N.

Один вариант – все (возможно, кроме одного) делители N известны. Делением N на все делители определяется и неизвестный последний делитель.

Другой вариант – неизвестны несколько делителей N. Остающиеся неизвестными делители могут быть определены делением исходного N на все известные делители. Если таким путем не все делители определяются, то задача сводится к факторизации составных меньших N чисел, определенных при предварительных делениях числа N на известные делители.

3. Аддитивная и мультипликативная формы представления чисел. Общий замысел нового подхода к задаче факторизации чисел состоит в следующем. В соответствии с теоремой факторизации этой процедуре подвергаются составные нечетные натуральные числа.

Аддитивная форма таких чисел – это сумма нечетного числа слагаемых, которые представляют собой непрерывный фрагмент последовательности нечетных чисел в НРЧ. Эта сумма, начиная с первого нечетного числа в ней, формируется далее нечетными числами, возрастающими всегда на две единицы. Для числа N могут существовать несколько различных сумм в разных областях (местах) НРЧ.

Пример 1. Пусть $N = 105$, тогда $N = 9+11+13+15+17+19+21 = 17+19+21+23+25 = 33 + 35 + 37 = 105$ представляется тремя различными нетривиальными суммами, а четвертая – тривиальная образована одним слагаемым N.

Далее, известно, что в НРЧ между квадратами последовательных чисел разности равны последовательным нечетным числам, которые могут описываться границами интервалов левой $\Gamma_л(N)$, и правой $\Gamma_п(N)$, соответствующих этим нечетным числам. Границы при этом всегда являются полными квадратами $N = \Gamma_п(N) - \Gamma_л(N)$.

Пример 2. Для сумм, представляющих $N = 105$, имеем (в кавычках записываются слагаемые сумм равные разностям указанных квадратов смежных чисел, имеющих разную четность):

4^2 «9» 5^2 «11» 6^2 «13» 7^2 «15» 8^2 «17» 9^2 «19» 10^2 «21» 11^2 или 8^2«17»9^2«19»10^2«21»11^2«23»12^2«25»13^2 или 16^2«33»17^2«35»18^2«37»19^2.

Мультипликативная форма чисел – это представление произведением разности квадратов границ интервала, соответствующего N и сформированного отрезком последовательности нечетных чисел. Если рассматривать суммы для $N = 105$, как интервалы (расстояния) между внешними границами крайних слагаемых в каждой из трех сумм примера 1, то получим представление числа $N = 105$ мультипликативной формой.

Пример 3. Выпишем мультипликативное представление квадратами: $11^2 – 4^2 = 13^2 – 8^2 = 19^2 – 16^2 = 105$ или в скобочном виде $105 =(11 – 4)\cdot(11 + 4) =7\cdot15 = (13 – 8)\cdot(13 + 8) = 5\cdot21= (19 – 16)\cdot(19 +16) = 3\cdot35$.

Представление N в такой форме как раз и обеспечивает разложение составного нечетного числа N на два нечетных сомножителя, из которых либо один, либо оба могут оказаться составными, либо, что случается более редко – оба простые числа.

Каждый полученный составной нечетный фактор можно подвергнуть далее такой же процедуре представления в аддитивной и мультипликативной форме.

Так необходимо действовать до получения в качестве всех факторов числа N только простых чисел.

Представляется, что рассмотренная алгоритмическая схема обработки составного нечетного числа N весьма слабо зависит от раз-

рядности N и сам процесс факторизации для больших, очень больших и малых чисел будет занимать практически одинаковое время, исчисляемое секундами или их долями при компьютерной обработке.

Таков общий замысел предлагаемого нового подхода к решению проблемы факторизации чисел. Реализация описанного процесса на практике встречает определенные трудности, преодоление которых возможно несколькими путями. Один из таких путей и рассматривается далее в работе. Необходимо осознать, что существующие на сегодняшний день практика и подходы к решению задачи факторизации больших чисел – это по существу путь «проб и ошибок», который ориентирован на использование свойств чисел, жестко зависящих от их разрядности. Чем больше разрядность числа, тем большее время требуется для его факторизации.

Существенное сокращение времени ожидается в алгоритмах факторизации, использующих свойства чисел, не зависящие от их разрядности. Такие свойства существуют и даже практически используются, например, признаки делимости чисел. Число N делится на три, если на три делится сумма (свертка) его цифр, доведенная до одной цифры. Факторизация таких чисел, описываемых сотнями и тысячами цифр, занимает секунды.

Пример 4. Пусть $N = 123456789$. Тогда сумма цифр фрагмента НРЧ $1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 = 45 \rightarrow 4 + 5 = 9 = 3 \cdot 3$.

Если число N имеет в записи сотни или даже тысячи цифр, то их сумма находится очень быстро и время ее вычисления слабо зависит от разрядности числа. Отсюда следует, что необходимо найти и использовать свойства чисел, не зависящие от их разрядности. Одно из таких свойств (ф-инвариант) числа используется для разработки алгоритма факторизации и позволяет преобразовать задачу факторизации в другую задачу – формирования разбиений специального вида для заданного числа N. Процедура такого преобразования (сведения) и рассматривается в предлагаемой работе.

Рассмотрим рисунок 1. На числовой оси x выделены четыре точки 2^2, 15^2, 110^2 и 111^2, в которых размещены квадраты натуральных чисел. Интервал между левой (первой) парой точек $15^2 - 2^2 = 221$ совпадает с интервалом между второй парой точек $111^2 - 110^2 = 221$. Если необходимо найти значения делителей меньшего $d_м$ и большего $d_б$ числа $N = 221$, то можно воспользоваться формулой сокращенных вычислений, а именно:

$$N = x_1^2 - x_o^2 = (x_1 - x_o)(x_1 + x_o) = d_м \cdot d_б,$$

где, в частности, для правой пары неизвестные переменные x_o и x_1 могут определяться выражениями:

$$x_o = (N - 1)/2 = 110,\ x_1 = (N + 1)/2 = 111,$$

и являются смежными натуральными числами, между квадратами которых лежит исследуемое число ***221.***

Подставляя вычисленные значения в формулу $N = (x_1 - x_o)(x_1 + x_o) = (111 - 110)(111 + 110) = 1 \cdot 221$, получаем делители меньший $d_м = 1$ и больший $d_б = 221$. Это тривиальное разложение, но, оказывается, существует еще пара квадратов, которая приводит к другому разложению числа $N = 221$ на множители, а именно:

$$N = (x_1 - x_o)(x_1 + x_o) = d_м \cdot d_б = (15 - 2)(15 + 2) = 13 \cdot 17,$$

которое является окончательным решением, так как оба делителя – простые числа.

Как видим, разрядность чисел нигде и никак себя не проявила. Проблема заключается в получении пар альтернативных чисел-квадратов разной четности, между которыми лежит факторизуемое число ***N***.

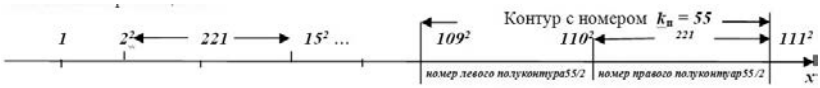


Рис. 1. Положение в НРЧ интервалов длиной $N = 221$ с квадратами в качестве границ

Тривиальный случай разложения здесь рассмотрен не зря. Именно он создает отправной посыл для разработки нового направления в решении проблемы факторизации больших чисел. Положение числа со значением $N = 221$ на числовой оси можно представлять отрезками НРЧ длиной из ***221*** позиции, среди которых только два отрезка будут с граничными позициями, содержащими полные квадраты. В результате изучения проблемы обнаружена такая характеристика нечетного числа (НЧ) и соответствующего ему интервала, которая является инвариантом интервала с границами-квадратами и длиной равной числу ***N*** независимо от того, где этот интервал на числовой оси размещается.

Известно, что любое натуральное нечетное число лежит между квадратами чисел. Тривиальное разложение на множители всегда позволяет указать пару квадратов для смежных чисел разной четности и числовую характеристику отрезка для числа, остающуюся неизменной при допустимых смещениях отрезка вдоль числовой оси. Допустимыми являются положения интервала длиной ***N*** , концы которого размещаются в точках квадратах натурального ряда чисел. Таких допустимых положений для интервала, соответствующего числу ***N***, тем больше, чем больше у ***N*** делителей. Сама эта характеристика представляет собой комбина-

торное разбиение со специальными свойствами некоторого другого (не ***N***) постоянного числа, зависящего от факторизуемого ***N***. В разбиении числа-константы, представляемого суммой меньших чисел, все слагаемые различаются на ***1*** кроме крайнего в сумме, от которого в сумму включается лишь половина слагаемого. Для рассматриваемого примера существует единственное допустимое смещение тривиального интервала длиной ***N = 221***, которое и изображено на рисунке 1. Оба этих интервала характеризуются одним разбиваемым числом-константой (ф-инвариантом), равным половине числа ***55.*** Ниже приводится представление константы в форме специального разбиения:

$$55/2 = 27.5 = ½ + 2 + 3 + 4 + 5 + 6 + 7.$$

4. Два специальных разбиения натурального числа *N*. Будем рассматривать задачу о разбиении натурального числа. Разбиением натурального числа ***N*** называется конечная невозрастающая последовательность натуральных чисел $k_0, k_1, k_2, k_3, \dots, k_t$, меньших ***N***, для которой $N = \sum_i^t k_i$, числа k_i называются блоками (частями) разбиения. Разбиения чисел бывают на нечетные и отдельно на четные части, а также упорядоченными и неупорядоченными [12]. Существуют разбиения чисел разбиения чисел на одинаковые и различные части и т.п.

Для наших целей будем использовать графическое представление *специальных разбиений* чисел на разные части. Специфика разбиений связана с ограничением на отличие ($\Delta = 1$) одной части разбиения от другой и в том, что крайняя часть (меньшая или большая) в сумме равна половине слагаемого, удовлетворяющего ограничению:

$$k_t - k_{(t-1)} = \dots = k_3 - k_2 = k_2 - k_1 = k_1 - k_0 = 1,$$

для чего воспользуемся точечным графом Феррера.

Далее будем использовать зависимости для целочисленных рядов [1]:

– сумма ***n*** натуральных чисел:

$$1 + 2 + 3 + \dots + n - 1 + n = n(n+1)/2 = C_{n+1}^2,$$

– сумма ***n*** натуральных чисел, где последнее слагаемое равно ***n/2***, (вариант специального разбиения):

$$1 + 2 + 3 + \dots + n - 1 + n/2 = n^2/2;$$

– сумма ***n*** нечетных натуральных чисел:

$$1 + 3 + 5 + \dots + (2n-3) + (2n-1) = n^2.$$

В задаче разбиения числа, связываемой с задачей факторизации, имеют дело с *двумя специальными случаями* разбиений не самого числа N, а его ф-инварианта, числа $k_п(N)/2$, т.е. половины номера предельного контура для N. Кроме того, будем различать разбиения числа $k_п(N)/2$, соответствующие целым и дробным числам, а также соответствующие *левым* ($N_л$) и *правым* ($N_п$) *натуральным нечетным числам*.

Графическое представление (см. таблица 2) множества разбиений чисел $k_п(N)/2$ имеет вид трапеции как составной части треугольной точечной диаграммы Феррера, образованной из ее подряд следующих строк. Такая трапеция вкладывается в треугольную диаграмму. Одно из оснований трапеции (верхнее или нижнее) всегда включает в сумму разбиения только половину своих точек.

Основная специализация рассматриваемых здесь двух типов разбиений заключается в том, что отличие частей разбиения числа одной от другой составляет лишь единицу ($\Delta = 1$), и только крайние части (строки основания трапеции графического представления) разбиения числа могут отличаться от соседних на величину большую, чем единица. Крайняя строка (верхняя для $N_л$, нижняя для $N_п$) всегда разбивается пополам. Если разбиваемая пополам строка содержит четное число точек, то все части разбиения числа N – целые числа, если число точек в такой строке нечетное, то крайняя часть разбиения – дробное число.

Отметим также некоторые другие особенности этих специальных разбиений. Эти особенности легко и наглядно воспринимаются при рассмотрении числовых примеров, которые и составляют основное содержание работы.

Все различные слагаемые в сумме разбиений ф-инварианта для числа N, кроме одного крайнего слагаемого, всегда отличаются на единицу, т.е. это числа k_i, следующие в натуральном ряде одно за другим:

$$k_п(N)/2 = \sum_{i=p}^{t} k_i \pm k/2,$$

где p – индекс номера начального (меньшего) контура;

t – индекс номера конечного (большего) контура;

$\pm$ – знак в сумме определяется видом (левое $k > k_t$, правое $k < k_p$) числа N;

k – номер крайнего контура (без индекса), от которого в сумме участвует только половина номера.

Тот факт, что сумма в специальном разбиении числа $k_п(N)/2$ формируется натуральными числами, возрастающими на 1, имеет ре-

шающее значение для нумерационной модели натурального числа. Все разбиения такого типа (все суммы $k_n(N)/2$) оказываются представленными в треугольной точечной диаграмме Феррера (графического представления разбиения числа) и могут быть получены из нее.

Во-первых, для N_n (правого нечетного числа) представление половины номера предельного контура (числа $k_n(N_л)/2$) разбиением всегда сформировано различными слагаемыми (частями), причем от от меньшего контура в сумму включается только половина его номера.

Во-вторых, для $N_л$ (левого нечетного числа) представление половины номера предельного контура (числа $k_n(N_л)/2$) разбиением всегда сформировано различными частями, если $k_n(N_л)/2$ – дробное число.

В-третьих, для $N_л$, если $k_n(N_л)/2$ – целое, то два слагаемых в сумме могут совпадать, причем, одно слагаемое из такой пары – это половина номера большего контура в сумме.

В-четвертых, все слагаемые во всех суммах – это интервалы позиций НРЧ, в которых крайние позиции заняты квадратами целых чисел. Поясним эти понятия числовыми примерами.

В средней части таблицы помещено графическое разбиение числа ***231*** на разные части (строки с точками), отличающиеся одна от другой на единицу. Для нечетных натуральных чисел N в пределах первой тысячи на основе этой диаграммы может быть проведена факторизация их на два фактора. Первые четыре колонки таблицы содержат характеристики интервальной модели (характеристики контуров). Колонка справа от точечной диаграммы – номера контуров. Две последние колонки – характеристики нумерационной модели. Уровнем названа n-я снизу строка таблицы. Значения характеристики в строках приведены с нарастанием от нижней строки вверх. В последнем столбце суммы точек из предшествующих строк суммируются с половиной количества точек текущей строки. В предпоследнем столбце суммы точек полных строк.

Для левых $N_л$ чисел такие суммы $k_n/2$, соответствующие интервалам длины $N_л$, всегда содержат определенное число точек и строк из треугольника и в верхней части лишь половину количества точек строки. Очевидно, если зафиксировать строку из суммы, которой соответствует значение равное $k_n/2$ или ближайшее большее, но также содержащее половину верхней строки, то для решения о номере нижней строки в сумме остается определить только номер нижнего (меньшего) контура или соответствующей строки, формирующей сумму $k_n/2$. Это значение определяется разностью между суммой точек для уровня фиксированной верхней строки и значением $k_n/2$. Если такая разность

присутствует в рассматриваемой колонке, то строка ей соответствующая, и строки ниже нее не учитываются в сумме.

Приведем таблицу 2 с такой диаграммой, сопроводив ее дополнительными сведениями об интервальной и нумерационной моделях числа ***N***.

Таблица 2**.** Характеристики интервальной и нумерационной моделей числа ***N***

Интервальная модель ***N***				Диаграмма Феррера	Нумерационная модель ***N***		
Правая граница контура	Средняя точка контура	Левая граница контура	Длина контура с №***k***	Графическое разбиение половины номера предельного контура $k_п$	Номер контура ***k***	Сумма точек C^2_{n+1} уровня	Значение $k^2/2$ для уровня
1849	1764	1681	168	ooooooooooooooooooooo	21	231=21·11	220.5
1681	1600	1521	160	oooooooooo|oooooooooo	20	210=21·10	200
1521	1444	1369	152	ooooooooooooooooo	19	190=19·10	180.5
1369	1296	1225	144	oooooooooo|oooooooooo	18	171=19·9	162
1225	1156	1089	136	ooooooooooooooooo	17	153=17·9	144.5
1089	1024	961	128	oooooooo|oooooooo	16	136=17·8	128
961	900	841	120	ooooooooooooooo	15	120=15·8	112.5
841	784	729	112	ooooooo|ooooooo	14	105=15·7	98
729	676	625	104	ooooooooooooo	13	91=13·7	84.5
625	576	529	96	oooooo|oooooo	12	78=13·6	72
529	484	441	88	ooooooooooo	11	66=11·6	60.5
441	400	361	80	ooooo|ooooo	10	55=11·5	50
361	324	289	72	oooooooooo	9	45=9·5	40.5
289	256	225	64	oooo|oooo	8	36=9·4	32
225	196	169	56	ooooooo	7	28=7·4	24.5
169	144	121	48	ooo|ooo	6	21=7·3	18
121	100	81	40	ooooo	5	15=5·3	12.5
81	64	49	32	oo|oo	4	10=5·2	8
49	36	25	24	ooo	3	6=3·2	4.5
25	16	9	16	oo	2	3=3·1	2
9	4	1	8	o	1	1=1·1	0.5

Поиск разности $C^2_{k+1} - k_п/2$ начинаем от значения $C^2_{k+1} > k_п/2$, а вычисленную разность сравниваем с $k^2/2$, до тех пор, пока они не совпадут. При несовпадении увеличиваем значение ***k***.

Пример 5. (*Возникновение равных слагаемых в специальном разбиении*). Задано СННЧ ***N = 119***, это число левое, так как сравнимо ***119 ≡ 3*(mod*4)*** с тройкой**.** Предельный контур для этого числа имеет длину ***L(119) = 119 + 121 = 240***. Номер предельного контура равен $k_п(119) = 240/8 = 30$**.** Значение ф-инварианта для числа ***119*** равно $k_п(119)/2 = 30/2 = 15$**.** Специальное разбиение этого ф-инварианта имеет вид ***15 = 3 + 4 + 5 + 6/2 = 3 + 4 + 5 + 3.*** В итоговой сумме все

слагаемые (кроме последнего) отличаются от соседних на $\Delta = 1$, и получились два одинаковых слагаемых, равных тройке. Чтобы убедиться, что разбиение ф-инварианта приводит к факторизации числа $N = 119$, восстановим N по разбиению, все слагаемые которого (кроме последнего) – это номера контуров. Последнее слагаемое 3 – номер полуконтура 6-го контура.

Умножаем слагаемые на $8: 3{\cdot}8 + 4{\cdot}8 + 5{\cdot}8 + 6{\cdot}8 = 24 + 32 + 40 + 48$. Последнее (большее равное 48) слагаемое должно давать в сумму только свой левый (меньший) полуконтур $23+25 = 48$, т.е. число 23. Итак, устанавливаем длину интервала для $N = 119$ (проверяем: $24+32+40+23 = 119$).

Остается вспомнить, что границами контуров и полуконтуров в НРЧ всегда являются квадраты целых чисел и получить их крайние значения у крайних слагаемых интервала:

– меньшая граница интервала $\Gamma_л(119) = \Gamma_л(3)$ – это левая граница для контура с номером 3,

$$\Gamma_л(3) = (2{\cdot}3 - 1)^2 = 25 = 5^2,$$

– большая граница интервала $\Gamma_п(119) = \Gamma_ц(6)$ – это правая граница для левого полуконтура с номером

$$6/2 = 3,\ \Gamma_ц(6) = (2{\cdot}6)^2 = 144 = 12^2.$$

Последний штрих: число $N = 119$ представляем расстоянием между границами интервала, т.е.

$$\Gamma_п(119) - \Gamma_л(119) = \Gamma_ц(6) - \Gamma_л(3) = 12^2 - 5^2 = (12 - 5)(12 + 5) = 7{\cdot}17 = 119$$

и получаем разложение N на множители.

5. Две модели натурального числа. Представим числовую ось, размеченную точками $x,\ x+1,\ x+2\ \ldots,$ пронумерованными числами натурального ряда, начиная от единицы. Вдоль этой оси перемещается движок с окошком (по типу логарифмической линейки), вмещающим N точек от точки с меньшим значением x_o, до точки с большим значением x_1 (натуральных чисел, нумерующих точки). Решение задачи факторизации можно представить следующим механизмом.

Перемещаем движок вдоль числовой оси так, чтобы крайние точки окна (x_1, x_o) в движке совпали с целыми квадратами разметки на оси. Такая ситуация достижима всегда для любых нечетных натуральных N, даже, если N простое число. Для ННЧ квадраты будут иметь разную четность, которая определяется для x_1 и x_o однозначно.

Для простого N задача решается исключительно просто, хотя в отличие от составных N, для простого числа существует лишь единственный вариант требуемого положения движка. Дело в том, что числа,

квадраты которых должны совпасть с крайними точками окна, всегда для нечетного простого числа соседние: одно четное, другое – нечетное $x_o = (N - 1)/2$, $x_1 = (N+ 1)/2$.

Действительно, квадраты этих чисел удовлетворяют задаче разложения на множители для любых N:

$$N = x_1^2 - x_o^2 = (x_1 - x_o)(x_1 + x_o) = 1 \cdot N.$$

Формальное решение задачи факторизации числа N получено, но, как следовало ожидать, оно тривиальное.

Для составных чисел N имеется два и/или более варианта положения движка на числовой оси, удовлетворяющих условиям задачи. Каждый вариант соответствует различным разложениям числа N на множители. Алгоритм решения задачи факторизации числа может обеспечивать получение одного нетривиального решения, после чего он, может быть применен, многократно повторяясь, но уже к найденным делителям (факторам), до полного разложения числа N на простые множители.

Таким образом, основная проблема заключается в нахождении нужного положения движка с «окном».

В основе сведения одной из названных в заголовке задачи к другой лежит принцип представления отдельного натурального числа моделью интервала числовой оси и представления моделью всего натурального ряда чисел нумерованными контурами. Из основной теоремы факторизации следует, что факторизации необходимо подвергать только отдельные трудно факторизуемые нечетные числа, поэтому далее кратко рассматриваются модели именно для таких чисел. Приведем некоторые понятия модели НРЧ.

Контуром (интервалом числа) в НРЧ называется непрерывное множество позиций, занимаемых последовательными натуральными числами, из которых меньшее является квадратом нечетного числа, а большее предшествует следующему по величине нечетному квадрату. Среди чисел контура обязательно присутствует один квадрат четного числа, лежащий между квадратами нечетных смежных чисел, формирующими границы контура. Длина (число позиций) любого контура кратна числу восемь. Таким образом, нечетные квадраты делят НРЧ на контуры, которые образованы всегда только двумя смежными ННЧ, разделяемыми четным квадратом.

Все контуры в НРЧ получают порядковые номера k, равные длине контура $L(k)/8$ поделенной на восемь.

Полуконтуром НРЧ называется часть контура (левая или правая), лежащая между квадратами чисел разной четности. Длины левого

и правого полуконтуров в одном контуре различаются на две единицы. Полуконтуры не снабжаются специальными номерами, но поскольку они нечетные числа, то для каждого из них легко определяется его порядковый номер в НРЧ. Пусть нечетное число $N = 2n - 1$, где $n = 0(1)\ldots$ – порядковый номер числа, тогда для любого натурального нечетного числа N его порядковый номер $n = (N + 1)/2$. Полуконтуру в модели НРЧ приписывается половина номера $k_п(N)/2$ его предельного контура, но не n.

Пример 6. Пусть ННЧ $N = 35$. Тогда $n = (35 + 1)/2 = 18$.

Интервалом числа НРЧ, соответствующим отдельному нечетному числу N, называется непрерывная последовательность контуров (полуконтуров), меньшей, чем N длины, суммарная длина которых равна N. Это означает, что любой интервал для полуконтура всегда начинается и заканчивается позициями квадратов чисел разной четности. Для отдельного составного числа в НРЧ, могут существовать несколько интервалов в разных частях НРЧ. Для простого числа такой интервал всегда один и он образован единственным полуконтуром *предельного контура* этого простого числа.

Интервальная модель натурального числа N представляет собой непрерывную последовательность контуров натурального ряда чисел, суммарная длина которых равна основной *характеристике* интервальной модели числа – значению факторизуемого числа N. В такой модели границами интервалов каждого нечетного числа N являются числовые квадраты разной четности. Это обстоятельство создает ряд неудобств, так как один из крайних контуров модели представлен в суммарной длине интервала лишь своей половиной (с границей – четным квадратом), т.е. полуконтуром.

При этом возникает два варианта положения такого полуконтура. Если факторизуемое нечетное число $N \equiv 3 (\mathbf{mod} 4)$, то граница интервала – четный квадрат является большим из двух (правая граница интервала для N), если число $N \equiv 1 (\mathbf{mod} 4)$, то четный квадрат меньший из двух (левая граница интервала для N).

Таким образом, вопрос о четности границ (правой, левой) для любого нечетного числа N решается однозначно, что, в свою очередь, определяет структуру интервальной модели числа N.

Структура интервала формируется конечным множеством непосредственно примыкающих друг к другу контуров и одного крайнего, примыкающего к ним полуконтура. Длина интервала определяется как сумма длин всех контуров и длины одного полуконтура. Эта модель далее преобразуется в другую модель путем замены длин контуров в сумме их номерами, а для крайнего полуконтура половиной номера

его предельного контура. После такой замены интервальная модель преобразуется в нумерационную модель, т.е. в сумму последовательных четных и нечетных натуральных чисел (номеров контуров). Неудобство создается и необходимостью учета для одного из крайних контуров лишь его полуконтура или половины номера контура.

Нумерационная модель натурального числа. Описанное ранее преобразование от интервальной модели к нумерационной модели числа упрощает ее, сводит к сумме последовательных натуральных чисел от некоторого значения $\boldsymbol{k}$ до значения $\boldsymbol{\ell}$. Замечательной особенностью этой модели является равенство этой суммы половине номера предельного контура $\boldsymbol{k_n(N)/2}$ нечетного числа $\boldsymbol{N}$. Независимо от положения в НРЧ исходного интервала числа $\boldsymbol{N}$ это равенство следует из теоремы о предельном контуре.

Значение $\boldsymbol{k_n(N)/2}$ находится элементарными простыми вычислениями для любого нечетного числа $\boldsymbol{N}$ и является основной *характеристикой* нумерационной модели числа. Для составных нечетных чисел $\boldsymbol{N}$ одно значение $\boldsymbol{k_n(N)/2}$ и суммы номеров контуров, образующих интервал для $\boldsymbol{N}$, может быть представлено разным количеством и разными по составу слагаемыми и, как следствие, разными разложениями числа $\boldsymbol{N}$ на два фактора.

Для нечетного числа $\boldsymbol{N}$ в НРЧ существует единственное представление, если число простое, и два или более представляющих интервалов, если число составное. Эти интервалы для составных чисел образуются контентом целых контуров и половиной одного крайнего из них. Интервал для $\boldsymbol{N}$ всегда образован нечетным числом полуконтуров, и это число является меньшим делителем $\boldsymbol{d_м}$ факторизуемого числа $\boldsymbol{N}$. Расположены контенты таких интервалов в разных частях НРЧ, на разном удалении от начала ряда. Особенностью интервалов является то, что граничными точками (числами) интервалов являются квадраты чисел разной четности. Среди таких интервалов всегда есть один, границами которого служат квадраты двух соседних чисел. Этот интервал называется *предельным полуконтуром*. Квадраты-границы альтернативных интервалов для числа $\boldsymbol{N}$ как бы раздвигаются по отношению к предельному. Количество полуконтуров, образующих интервал, тем больше, чем ближе интервал к началу НРЧ, так как контуры имеют меньшую длину. Заметим, что длина среднего полуконтура (при их нечетном количестве) равна $\boldsymbol{d_б}$ большему делителю числа $\boldsymbol{N}$.

Длина предельного интервала, как и всех других альтернативных, равна значению числа $\boldsymbol{N}$, но сам интервал при этом представляет собой лишь половину контура, который также называется *предельным контуром*.

Номер предельного контура важнейшая характеристика нечетного числа N. Он обозначается символом $k_п$ и вычисляется по формуле:

$$k_п = L_n(N) /8,\ L_n(N) = f(N) = N_л + N_п.$$

Здесь $N_л$ и $N_п$ это полуконтуры предельного контура, а число N может быть любым из них. Четным квадратом в предельном (и в любом другом) контуре (общая граница полуконтуров) является квадрат его удвоенного номера $(2k_п)^2$. Тогда границы контура и значения его полуконтуров определяются через его номер $k_п$, как:

$$Г_п(N) = (2k_п + 1)^2,\ Г_л(N) = (2k_п - 1)^2 \text{ и } N_i = 4k_п \pm 1.$$

Знак в последнем выражении выбирается в зависимости от класса (левый, правый) нечетного числа N. Длина предельного контура определяется как разность его границ:

$$L_п(N) = Г_п(N) - Г_л(N) = 8k_п.$$

Границы всех других (альтернативных) интервалов образуют квадраты несмежных натуральных чисел, но также разной четности и в том же порядке.

6. Заключение. Рассмотренные в работе вопросы позволяют сделать некоторые выводы об описываемых задачах и о проблеме факторизации в целом. На факторизацию как на проблему явно не указывали великие математики, а некоторые из них лишь косвенно ее затрагивали Диофант, Ферма, Гаусс, Эйлер, Гильберт, не включая в перечень нерешенных задач. Возможно, именно это и притормаживало развитие теории в этом направлении, пока острой потребности в арифметической операции обращения умножения не возникало.

Развитие теории криптологии (двухключевые системы) всколыхнуло математическую мысль, но ни компьютерная вооруженность, ни распределенные сетевые вычисления к быстрому успеху не привели. Отсутствие фундаментальных теоретических результатов о таком объекте как НРЧ не позволяет найти выход из возникшего математического тупика. Даже финансовая стимуляция исследований не привела к ускорению процесса решения ЗФБЧ.

В работе предложена общая схема обработки составного натурального числа, базирующаяся на основной теореме арифметики и теореме факторизации, приводящая к разложению числа на множители и исключающая перебор вариантов. Рассмотренный подход к решению ЗФБЧ основывается на использовании закономерностей структурного

построения НРЧ и свойства натуральных чисел, не зависящего от их разрядности.

Литература

1. *Бронштейн И.Н., Семендяев К.А.* Справочник по математике для инженеров и учащихся ВТУЗов // М.: ГИТТЛ. 1954. 608 с.
2. *Василенко О.Н.* Теоретико-числовые алгоритмы в криптографии // М.: МЦНМО. 2003. 328 с.
3. *Ваулин А.Е.* и др. Фундаментальные структуры натурального ряда чисел // Сб.тр. 7-го Международного симпозиума. М.: РУСАКИ. 2006. С. 384–387.
4. *Ваулин А.Е.* Новый метод факторизации больших чисел в задачах анализа и синтеза двухключевых криптографических алгоритмов. Ч.1. // Информация и космос. 2005. №3. С. 74–78.
5. *Ваулин А.Е.* Новый метод факторизации больших чисел в задачах анализа и синтеза двухключевых криптографических алгоритмов. Ч.2. // Информация и космос. 2005. №4. С. 104 –112с.
6. *Дэвенпорт Г.* Высшая арифметика // М.: Наука. 1966. 176 с.
7. *Евклид*. Начала. М–Л. 1948–1950. Т. 1–3.
8. RSA. URL: https://ru.wikipedia.org/wiki/RSA.
9. *Ноден П., Китте К.* Алгебраическая алгоритмика (с упражнениями и решениями) // М.:Мир. 1999. 720 с.
10. *Пойя Д.* Математика и правдоподобные рассуждения // М.: ИЛ. 1957. 464 с.
11. *Ферма П.* Исследования по теории чисел и диофантову анализу // М.: Наука. 1992. 320 с.
12. *Эндрюс Г.* Теория разбиений // М.: Наука. 1982. 256 с.

References

1. Bronshtejn I.N., Semendyaev K.A. *Spravochnik po matematike dlja inzhenerov i uchashhihsja VTUZov* [Handbook of mathematics for engineers and students VTUZov]. M.: GITTL. 1954. 608 p. (In Russ.).
2. Vasilenko O.N. *Teoretiko-chislovye algoritmy v kriptografii* [Number-theoretic algorithms in the cryptography]. M.: MTsNMO. 2003. 328 p. (In Russ.).
3. Vaulin A.E. et al. [The fundamental structure of the naturally row numbers]. *Sb.tr. 7-go Mezhdunarodnogo simpoziuma*. [Proceedings of the 7th International Symposium]. M.: RUSAL KI. 2006. pp. 384–387. (In Russ.).
4. Vaulin A.E. [A new method of factoring large numbers in the analysis and synthesis of two-key cryptographic algorithms. Part 1]. *Informacija i kosmos – Information and Space.* 2005. no. 3. pp. 74– 78. (In Russ.).
5. Vaulin A.E. [A new method of factoring large numbers in the analysis and synthesis of two-key cryptographic algoritmov. Part 2]. *Informacija i kosmos – Information and Space.*. 2005. no. 4. pp. 104–112. (In Russ.).
6. Davenport G. *Vysshaja arifmetika* [Higher Arithmetic]. M.: Nauka. 1966. 176 p.
7. Euclid. *Euclid's Elements*. M-L. 1948–1950. vol. 1–3. (In Russ.).
8. RSA. Available at: https://ru.wikipedia.org/wiki/RSA. (In Russ.).
9. Noden P., Kitte K. *Algebraicheskaja algoritmika (s uprazhnenijami i reshenijami)* [Algebraic algorithmics (with exercisingtions and decisions)]. Moscow: Mir, 1999. 720 p. (In Russ.).
10. Pojja D. *Matematika i pravdopodobnye rassuzhdenija* [Mathematics and plausible reasoning]. M.: IL, 1957. 464 p. (In Russ.).
11. Ferma P. *Issledovanija po teorii chisel i diofantovu analizu* [Studies in number theory and diophantine analease]. M.: Nauka. 1992. 320 p. (In Russ.).

12. Andrews G. *Teorija razbienij* [The Theory of partitions]. M.: Nauka. 1982. 256 p. (In Russ.).

Ваулин Арис Ефимович — к-т техн. наук, доцент, доцент кафедры систем сбора и обработки информации, Военно-космическая академия имени А.Ф. Можайского. Область научных интересов: криптоанализ, теория автоматов. Число научных публикаций — 200. yourmail_@mail.ru; ул. Ждановская д.13, Санкт-Петербург, 197082; р.т.: +7(812)347-9687.

Vaulin Aris Efimovich — Ph.D., associate professor, associate professor of system for collecting and processing information department, Mozhaisky military space Academy. Research interests: information security in automated systems for special purposes, cryptanalyst. The number of publications — 200. yourmail_@mail.ru; 13, Zhdanovskaya street, St. Petersburg, 197198, Russia; office phone: +7(812)347-9687.

Назаров Михаил Сергеевич — адъюнкт кафедры систем сбора и обработки информации, Военно-космическая академия имени А.Ф. Можайского. Область научных интересов: схемотехника, микроэлектроника. Число научных публикаций — 10. mikl21@mail.ru; ул. Ждановская д.13, Санкт-Петербург, 197082; р.т.: +7(812)347-9687.

Nazarov Mikhail Sergeevich — adjunct of systems for collecting and processing information department, Mozhaisky Military Space Academy. Research interests: circuitry, microelectronics. The number of publications — 10. mikl21@mail.ru; 13, Zhdanovskaya street, St. Petersburg, 197082, Russia; office phone: +7(812)347-9687.

РЕФЕРАТ

Ваулин А.Е., Назаров М.С. **Сведение задачи факторизации натурального числа к задаче разбиения числа на части. Часть 1.**

Развитие механизмов факторизации составных целых чисел рассматривается работе. От современных методов не следует ожидать, что алгоритм будет быстродействующим и эффективным в ближайшее десятилетие, в связи с ограниченным подходом к учету свойств чисел математический подход к решению этой проблемы, который базируется на методах типа решета Эратосфена, не является перспективным.

Механизм, предлагаемый автором этой работы, использует совершенно новый подход, основанный на изучении внутреннего строения натурального ряда и применения ряда свойств, слабо зависящих от разрядности числа. Примером такого свойства является признак делимости числа на 3. Такой подход обеспечивает переход от факторизации целых чисел к поиску специального свойства названного ф-инвариантом числа. При этом ожидается, что проблема будет менее сложной.

SUMMARY

Vaulin A.E., Nazarov M.C. **Conversion of Integer Factorization to a Problem of Decomposition of a Number. Part 1**.

The development of factorization mechanisms of composite integer numbers is being examined in this work. The current methods should not be expected to become more rapid and efficient in the nearest decade, due to narrow and inadequate mathematical approach to solution of this problem, which is based on so-called sieve of Eratosthenes.

The mechanism suggested by author of this work, uses a completely new method, based on examination of internal structure of natural sequence and application of digit place independent features (the criterion for divisibility). That kind of approach provides a conversion from integer factorization to a retrieval of the special figure separation, so-called F-invariant, which turns out to be less complex problem.

УДК 621.382

А.А. Мусаев

АДАПТИВНАЯ МУЛЬТИРЕГРЕССИОННАЯ ОЦЕНКА В УСЛОВИЯХ ХАОТИЧЕСКИХ ПРОЦЕССОВ ВАЛЮТНОГО РЫНКА

Мусаев А.А. **Адаптивная мультирегрессионная оценка в условиях хаотических процессов валютного рынка.**

Аннотация. Рассмотрена задача мультирегрессионной оценки стоимости валютного инструмента на основе адаптивного выбора регрессоров, образованных группой валютных пар, наиболее коррелированных с оцениваемым активом. В условиях хаотической динамики котировок валютных инструментов степень корреляции между валютными парами изменяется во времени. Отсюда возникает задача адаптивного оценивания с переменным составом группы регрессоров. Для оценки потенциального выигрыша, достигаемого при использовании управляющей стратегии на основе предложенного подхода, используется метод эволюционного моделирования.

Ключевые слова: хаотические процессы, мультирегрессионная оценка, корреляционный анализ, численный анализ, адаптация, эволюционное моделирование, валютный рынок, валютные инструменты, Forex.

Musaev A.A. **Adaptive Multiregression Currency Estimation in the Chaotic Market Environment.**

Abstract. The problem of multiregression estimation of the currency cost is considered. The offered approach is based on an adaptive choice of the regressors formed by group of currency pairs, the most correlated with an estimated asset. In the conditions of chaotic dynamics of currency quotations, correlation degree between currency pairs changes in time. From here the problem of adaptive estimation with variable structure of group of regressors follows. The method of evolutionary modeling is used for an assessment of the potential prize, reached when using the corresponding control strategy

Keywords: chaotic processes, multiregression estimation, correlation analysis, numerical analysis, adaptation, evolution modeling, currency, Forex.

1. Введение. Динамика котировок на рынках капитала носит хаотический характер [1–4]. Символом хаоса является его непредсказуемость. Однако существуют закономерности столь высокого порядка, что им подчиняется даже динамика хаоса – законы диалектики [5]. В соответствии с гегелевскими законами происходит отрицание отрицания хаоса. Иными словами, хаос неизбежно порождает порядок. Умение обнаружить и использовать локальные проявления порядка являются характерной чертой трейдеров-профессионалов, отличающих от огромного коллектива неудачников валютного рынка.

Важным направлением поиска упорядоченных структур в хаосе является анализ взаимных корреляционных связей, т.е. переход в область многомерного анализа данных [6, 7]. Рассмотрение поведения конкретного рыночного актива на фоне динамики тесно связанного с ним сегмента валютного рынка создает предпосылку для построения

управляющих стратегий на основе использования локальных упорядоченных структур – коррелированных групп наблюдения. Действительно, наличие локальных корреляционных связей позволяет формировать скользящие оценки рыночной стоимости валютного актива по совокупности текущих наблюдений котировок связанных с ним валютных инструментов.

Текущее значение котировок может существенно отличаться от ее рыночной оценки, т.е. оценки, отражающей рыночные представления о ее стоимости. Это связано с наличием статистических флуктуаций, обусловленных большим числом полностью или частично неконтролируемых факторов. В этом случае рынок будет стремиться устранить данное несоответствие, что неизбежно предопределит направление движения котировки актива. Указанное свойство служит основой для построения так называемых осцилляторов [8], т.е. индикаторов состояния рынка, основанных на недооценке или переоценке текущей стоимости актива.

Пример построения *мультирегрессионной* (MR, multiregression) оценки, основанной на совокупности регрессоров из трех ведущих валютных пар (EURUSD, EURJPY, USDJPY) и исследование ее эффективности приведен в [9]. В этой же работе приведены примеры, иллюстрирующие влияние размера окна наблюдения на качество MR оценки в условиях рыночной динамики котировок. Очевидно, что механистический выбор регрессоров, не учитывающий реального уровня их корреляционной связи с активом, существенно снижает эффективность соответствующей управляющей стратегии. В связи с этим в настоящей работе рассмотрена задача оценки текущей стоимости актива на основе группы валютных пар, наиболее коррелированных с указанным валютным активом.

Важно заметить, что корреляционная матрица хаотического многомерного процесса является нестационарной и неэргодичной. Однако ее изменения, в отличие от исходных процессов динамики котировок, как показано в [7], достаточно инерционны. Это позволяет перейти к адаптивной схеме, основанной на периодическом пересчете корреляционной матрицы валютного рынка и, при необходимости, к изменению состава группы регрессоров.

Другое важное замечание связано с математической некорректностью применения алгоритмов обработки данных, основанных на вероятностно-статистической парадигме, к хаотическим процессам. В частности, нарушается базовый постулат вероятностной аксиоматики, предполагающий повторяемость опытов в идентичных условиях. В свою очередь, некорректность применения статистических алгоритмов

оценивания неизбежно приводит к снижению их эффективности. Реализация аналитических исследований качества регрессионных оценок потребует введения ограничений, не отвечающих характеру протекания хаотического процесса. Поэтому установление границ допустимости (точнее – пригодности) применения алгоритмов статистической обработки возможно лишь на основе экспериментального численного анализа.

Последнее замечание связано с необходимостью оценки потенциальной эффективности управляющей стратегии, основанной на адаптивном MR-оценивании. В настоящей работе такая оценка, осуществлялась методом эволюционного моделирования [10].

2. Постановка задачи. Пусть $\{Y_k,\ \ k=1,...n\}$ - дискретный временной ряд наблюдений текущей стоимости валютного инструмента. Стоимость актива подвержена случайным флуктуациям, приводящим к возможным отклонениями ее от представления рынка об ее истинной стоимости. В качестве оценки рыночной стоимости используется регрессионная оценка вида:

$$\hat{Y}_k = \sum_{j=1}^{m} c_j X_{kj},\quad k=1,...,n,$$

где X_{kj} - значения группы из m регрессоров на k-й момент времени наблюдения, образующие матрицу наблюдений $X=\{X_{kj},\ \ k=1,..n,\ \ j=1,...,m\}$, $c=(c_1,c_2,...,c_m)^T$ - вектор коэффициентов регрессии. Наличие оперативной оценки текущей стоимости валютного инструмента $\{\hat{Y}_k,\ \ k=1,...n\}$ позволяет построить вариант управляющей стратегии, основанной на величине и знаке разности $\{d_k=\hat{Y}_k-Y_k,\ \ k=1,...n\}$.

В дальнейшем для описания многомерной линейной регрессии будем использовать матричную нотацию [11, 12] вида $\hat{Y}=Xc$.

Минимизируя сумму квадратов ошибок $\varepsilon=(\varepsilon_1,\varepsilon_2,...,\varepsilon_m)^T$

$$\sum_{i=1}^{n}\varepsilon^2=\varepsilon^T\varepsilon=(Y-Xc)^T(Y-Xc),$$

приходим к системе нормальных уравнений, решение которых, в свою очередь, позволяет определить хорошо известное соотношение для оценки параметров регрессии по МНК (*метод наименьших квадратов*) $\hat{c}=(X^TX)^{-1}X^TY$.

Для оценки векторного коэффициента передачи $\hat{c}$ будет использоваться скользящее окно наблюдения:

$$w_i = (Y_{i-m}, Y_{i-m+1}, ..., Y_i), \quad i = (m+1), ..., n,$$

позволяющее снизить влияние нестационарности на качество обработки данных. Организация скользящего окна наблюдения и отвечающая ему схема структуризации данных представлена в [13]. При этом информационная платформа анализа формируется в виде двумерной таблицы, в левой части которой представлены данные наблюдений за регрессорами, в роли которых, в данном случае, выступают валютные инструменты рынка Forex. Правая часть таблицы представлена наблюдениями за оцениваемым активом.

Величину разности $d_i = \hat{Y}_i - Y_i$, где $\hat{Y}_i = \hat{c}_i X_i$ - оценка стоимости актива, сформированная на основе текущих значений регрессоров, можно использовать для анализа текущего состояния наблюдаемого валютного инструмента.

В силу хаотичности, а следовательно, и нестационарности ряда наблюдений, корреляционная матрица валютных пар $R_{ij} = cor(Y_i, Y_j)$ изменяется во времени. Следовательно, оптимальной состав фиксированной по размеру группы регрессоров $\{ X_{kj}, \quad k = 1,...n, \quad j = 1,...,m \}$ также будет меняться и его необходимо корректировать. В связи с этим используется адаптивная схема регрессионного оценивания с периодической коррекцией состава указанной группы регрессоров.

В качестве модели валютного рынка используем группу из 16-ти наиболее часто используемых валютных инструментов, представленных в табл. 1. В дальнейшем для обозначения валютных пар будем использовать их номера в этой таблице.

Таблица 1. Валютные инструменты

№№	1	2	3	4
Инструмент	EURUSD	EURJPY	EURGBP	EURCHF
№№	5	6	7	8
Инструмент	EURCAD	USDCAD	USDCHF	USDJPY
№№	9	10	11	12
Инструмент	GBPCHF	GBPJPY	GBPUSD	AUDJPY
№№	13	14	15	16
Инструмент	AUDUSD	CHFJPY	NZDUSD	NZDJPY

Для оценки коэффициентов корреляции между валютными парами будем использовать известное соотношение $r_{ij} = s_{ij} / \sqrt{s_{ii} s_{jj}}$, где

s_{ij}, $i,j=1,...,m$ - коэффициенты ковариации, образующие в совокупности матрицу ковариаций:

$$S = X^T X /(n-1) = \{ s_{ij}, \quad i,j=1,...,m \}.$$

В качестве оптимальной группы регрессоров выбираются m валютных пар, обладающих наибольшими значениями коэффициента корреляции с валютной парой, используемой в качестве рабочего актива для получения спекулятивного выигрыша.

3. Предварительный анализ корреляций. В качестве предварительной задачи рассмотрим оценку матрицы корреляций $R = \{ r_{ij}, \quad i,j=1,...,m \}$ для совокупности наблюдений за всеми 16 параметрами на временном интервале в 100 дней.

Тональное представление матриц корреляций между 16 валютными инструментами приведено на рисунке 1. Наиболее светлые тона соответствуют сильной положительной корреляционной связи, а наиболее темные – отрицательной.

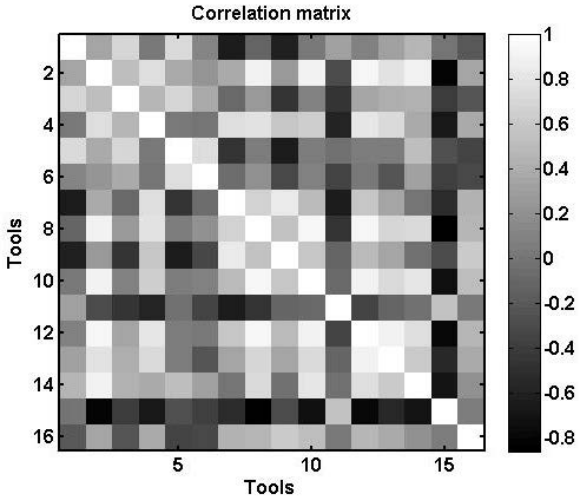


Рис. 1. Тональное представление матриц корреляций между 16 валютными инструментами

Практический вывод из рассмотрения тональной матрицы состоит в существенном разбросе значений коэффициента корреляции для различных валютных пар. Достаточно очевидно, что если коэффициент валютной пары относительно используемого актива колеблется

в пределах (-0.5, +0.5),то такую пару не следует использовать в качестве регрессора, ее значения практически не содержат полезной информации, используемой для оценивания актива.

Как уже отмечалось выше, динамика котировок является хаотическим процессом, а следовательно, и нестационарным. Это приводит к изменению значений корреляционных связей во времени. При этом данное изменение происходит относительно медленно и нет необходимости пересчитывать значения корреляционной матрицы на каждом шаге наблюдений.

В качестве грубого приближения будем использовать период между такими пересчетами в 8-10 часов. Об оптимальности выбора такого интервала говорить сложно, поскольку, в силу хаотичности наблюдаемых процессов, любая оптимальная совокупность параметров будет условной и привязанной к определенному временному интервалу ретроспективных наблюдений. Смена интервала наблюдений (даже при очень больших размерах наблюдений) неизбежно приведет к изменению значений оптимальных параметров. Такова природа хаоса.

4. Пример. Реализация простейшей управляющей стратегии с адаптивной MP оценкой. Рассмотрим в качестве примера изменения группы оптимальных регрессоров для валютного актива, образованного седьмым валютным инструментом USDCHF. На рис. 2 представлена динамика изменения котировки в течение 10 дней на фоне изменения котировок пяти валютных пар, наиболее коррелированных с ним на указанном интервале наблюдения. Соответствующая группа образована валютными парами с номерами [1, 8-10, 16].

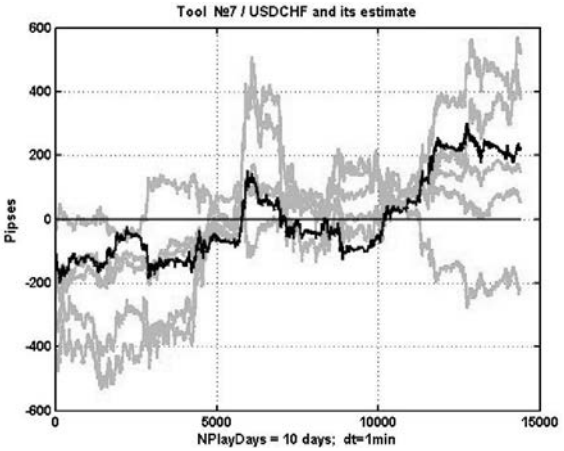


Рис. 2. Динамика изменения котировки валютного инструмента и группы 5 наиболее коррелированных с ним валютных пар

Заметим, что один из элементов ведет себя противофазно по отношению к изучаемому процессу. Это связано с тем, что степень взаимосвязи оценивается по модулю. Валютная пара с сильной отрицательной связью также несет в себе большой объем информации о поведении связанного с ним инструмента. При этом соответствующий регрессионный коэффициент перед этим членом будем иметь отрицательный знак.

В рамках принятых ограничений, будем осуществлять пересчет корреляционной матрицы рынка (т.е. всех 16-ти валютных пар). Из полученной матрицы выбирается строка, соответствующая номеру рабочего актива, и упорядочивается по убыванию значений модулей. Наблюдения полученного вариационного ряда со 2-го по (m+1)-й определяют группу регрессоров с наибольшими по модулю значениями корреляционных связей. Соответствующие результаты, полученные для 24-х непересекающихся 10-часовых интервалов наблюдения, представлены в таблице 2.

Таблица 2. Упорядоченные по убыванию степени коррелированности списки регрессоров на непересекающихся интервалах наблюдений длительностью 10 часов

Интервалы наблюдения	Номеров регрессоров
1-7	1 9 8 10 16
8-10	9 1 10 8 16
11	9 10 1 8 16
12-14	9 8 10 16 1
15-17	9 8 10 16 1
18-19	8 9 10 16 4
20-21	8 9 10 4 16
22-24	9 8 10 4 16

Из приведенных данных видно, что в течение первых семи интервалов наблюдения оптимальная группа регрессоров <1 9 8 10 16> не менялась. На 8-10 шагах состав группы также сохранился, но первый и девятый регрессоры поменялись местами. Дальнейшая эволюция состава группы регрессоров понятна из приведенных в таблице данных.

Общий вывод состоит в том, что состав группы регрессоров меняется достаточно медленно и 10-дневный интервал адаптации является вполне приемлемым для формирования регрессионных оценок с заданным составом регрессоров.

На рисунке 3 представлен пример реализации простейшей управляющей стратегии, основанной на адаптивной MR-оценке. В случае, если значения разности d_k между оценкой и текущей стоимо-

стью валютного инструмента оказывается больше (по модулю) порогового значения B, формируется рекомендация на открытие позиции в соответствующую сторону.

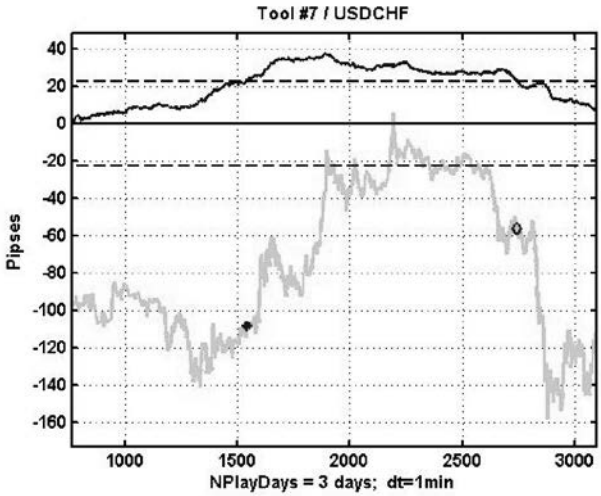


Рис. 3. Пример реализации простейшей управляющей стратегии с адаптивным MP осциллятором

На рисунке 3приведен центрированный график динамики котировки валютного инструмента USDCHF (здесь - нижний график) и сглаженный график d_k (верхний график) на интервале наблюдения в 3 дня. Звездочкой отмечено состояние котировки в момент открытия позиции вверх. Ромбик соответствует моменту закрытия позиции, осуществляемому при обратном пересечении графика d_k порогового значения B.

5. Оценка потенциальных характеристик управляющей адаптивного MP осциллятора на основе эволюционной оптимизации параметров стратегии. Отсутствие аналитического представления исходного хаотического процесса не позволяет получить точную оценку потенциальных возможностей выбранной управляющей стратегии. Наиболее эффективным способом получения такой оценки является численный анализ, основанный на случайном поиске оптимальных параметров управляющей стратегии. Вариантом реализации случайного поиска является технология эволюционного моделирования, предложенная в [10], и нашедшая широкое применение в задачах численной оптимизации [14–19]. Особенности применения эволюционно-

го моделирования в задачах параметрической оптимизации управляющих стратегий приведены в [20].

В качестве примера рассмотрена задача оптимизации описанной выше простейшей управляющей стратегии, когда открытие позиции осуществляется при пересечении сглаженного значения $\{ d_k = \hat{Y}_k - Y_k,\ \ k = 1,...n \}$ порогового значения $\pm B$. Закрытие позиции производится при обратном пересечении процессом d_k уровня $\pm B/2$.

Список оптимизируемых параметров стратегии $G = [nW, \alpha, B]$ (в терминах эволюционного моделирования – геном G) включает в себя размер скользящего окна наблюдения nW, на котором производится регрессионная оценка, коэффициент экспоненциального сглаживания α и уровень принятия решения B.

На начальном этапе путем внесения малых (в пределах среднеквадратических отклонений (ско) соответствующих параметров) вариаций во все параметры формируется группа геномов-предков или анцесторов (ГА) размером N_a. Далее, в цикле по числу поколений N_{gc}, формируется новое поколение, состоящее из уже сформированной группы геномов-предков и формируемой группы геномов-потомков или дескендеров (ГД). Геномы потомки формируются из геномов-предков тремя основными способами [20], включающими в себя:

1. Небольшие единичные изменения, вносимые в один из параметров ГА. Выбор параметра осуществляется случайным розыгрышем. Если же предполагается вносить изменения последовательно в каждый параметр, то каждый ГА получает m_g модификаций, где m_g - размер генома. В этом случае возникает $N_d^{(1)} = N_a m_g$ потомков с заданным типом модификации, причем в каждом из них модифицируется только один параметр (ген). В данном случае $m_g = 3$, следовательно, если в каждом поколении сохранять $N_a = 4$ наилучших вариантов (предков), получим $N_d^{(1)} = 12$ версий ГД первого типа.

2. Небольшие групповые изменения. Осуществляется аналогично *SSM,* но изменения вносятся не в один, а сразу во все параметры. Таким образом, возникает еще $N_d^{(2)} = 4$ версии ГД с медленными изменениями во всех генах.

3. Сильные единичные изменения или параметрическая мутация. Выбор ГА и номера гена осуществляется случайным розыгрышем. С вероятностью параметрической мутации P_{pm} получает $N_d^{(3)}$ потомков, в каждом из которых модифицируется один ген в диапазоне $|\Delta| > 3\sigma$.

В качестве примера использовалась программа с числом смены поколений $N_{gc} = 9$ на одном и том же временном интервале в 10 дней. В качестве начального генома использовался вектор $G_0 = [nW_0, \alpha_0, B_0] = [5, 0.01, 0.6]$. При формировании модифицированных геномов использовались грубые оценки ско трех перечисленных параметров $SkoG = [3, 0.02, 0.5]$.

Для сравнения потенциальной эффективности управляющих стратегий, основанных на МР оценке состояния используемого актива рассматривались два варианта:

1. Вариант с фиксированной группой из пяти регрессоров, выбранных перед началом торговых операций на основе критерия максимальной коррелированности с рабочим инструментом (активом). Корреляционная матрица для 16 финансовых инструментов оценивалось на основе наблюдений за их котировками в течение 15 дней, предшествующих началу торгов.

2. Вариант с последовательной коррекцией группы из пяти регрессоров. Коррекция осуществлялась на основе того же критерия максимальной коррелированности с рабочим инструментом (активом) с интервалом в 10 часов. Оценка корреляционной матрицы осуществлялась по результатам наблюдений за их котировками на скользящем окне наблюдения размером также в 15 дней.

Поскольку оценка выигрыша осуществлялась с использованием случайного поиска, можно говорить лишь о некотором приближении к оптимальному решению, которое теоретически можно было бы получить путем полного перебора значений параметров управляющей стратегии.

На рисунке 4 представлен пример реализации наилучшего варианта параметров управляющей стратегии, полученный в результате эволюционной параметрической оптимизации в течение 9 поколений соответствующих программ. Описание приведенных графиков аналогично описанию выше представленных графиков на рисунке 3. Данный график соответствует первому варианту, т.е. неадаптивному варианту по отношению к выбору набора регрессоров. Слева от указанного

графика, на рисунке 5 приведена зависимость роста выигрыша в зависимости от номера поколения для неадаптивной стратегии.

Аналогичные графики реализации субоптимальной стратегии и зависимости выигрыша от номера поколения для второго варианта, основанного на последовательной коррекции списка регрессоров, приведены, соответственно на рисунках 6 и 7.

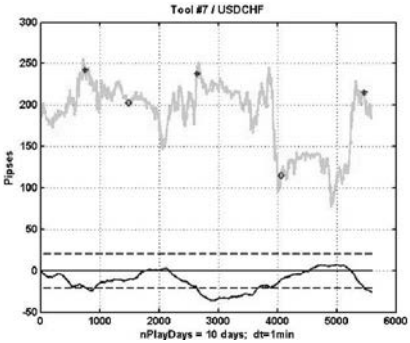


Рис. 4. Пример реализации субоптимальной управляющей стратегии с неадаптивным MP осциллятором

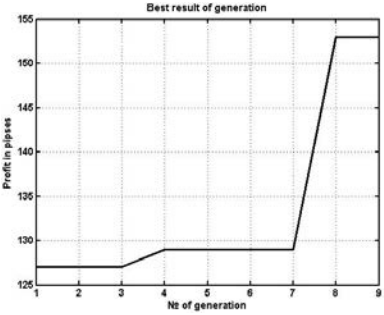


Рис. 5. Зависимость роста выигрыша в зависимости от номера поколения для неадаптивной стратегии

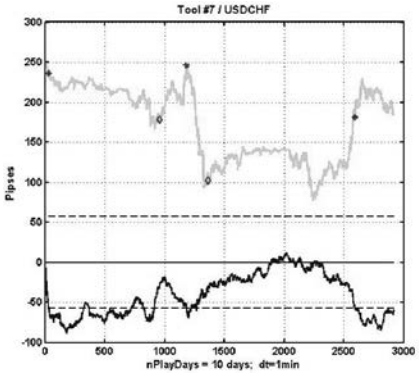


Рис. 6. Пример реализации субоптимальной управляющей стратегии с адаптивным MP осциллятором

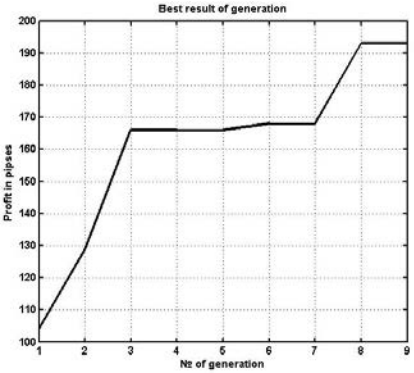


Рис. 7. Зависимость роста выигрыша в зависимости от номера поколения для адаптивной стратегии

Сравнение двух приведенных примеров показывает, что наличие адаптации при выборе группы регрессоров позволяет повысить качество регрессионной оценки и, как следствие, повышает уровень потенциального выигрыша примерно на 27%. Разумеется, отдельный

пример не дает объективной картины выигрыша. Для объективизации результата использовалось усреднение по 100 10-дневным участкам наблюдения котировок.

Заметим, что усреднение по реализациям в данном случае не эквивалентно усреднению по одному участку длиной, равной сумме отдельных реализаций. Это связано с тем, что хаотический процесс динамики котировок не является эргодичным. Поэтому задача воспроизводилась для обеих схем усреднения и показала, что выигрыш от адаптации колеблется в диапазоне 20-25%.

Заметим, что представленные в настоящей работе исследования указывают лишь на возможность повышения качества МР оценки в условиях хаотической динамики. В то же время основанная на ней управляющая стратегия представляет собой традиционный осциллятор со всеми присущими ему достоинствами и недостатками, описанными, например, в [8, 9].

6. Заключение. Оценка эффективности адаптационных технологий в задачах построения управляющих стратегий, ориентированных на функционирования в условиях хаотической динамики, является неоднозначной. Это связано с тем, что хаос, в силу своей нестационарности и неэргодичности, не позволяет замкнуть контур адаптации настолько быстро, насколько меняется структура наблюдаемого динамического процесса.

Однако в качестве вспомогательного инструмента, ориентированного на параметры с относительно медленными изменениями, адаптация может оказаться вполне полезной. В частности, как показано в настоящей работе, адаптация к вариациям корреляционной структуры многомерной динамики котировок может повысить качество восстановления стоимости актива, и как следствие, поднять уровень потенциальной эффективности мультирегрессионного осциллятора.

Литература

1. *Peters E. E.* Chaos and order in the capital markets: a new view of cycles, prices, and market volatility (2nd ed.). NY: John Wiley & Sons. 1996. 288 p.
2. Williams B.M. Trading chaos // NY: John Wiley & Sons. 2002. 251 p.
3. *Мусаев А.А.* Моделирование котировок торговых активов // Труды СПИИРАН. 2011. Вып. 17. С. 5–32.
4. *Колодко Д.В.* Нестационарность и самоподобие валютного рынка Forex // Управление экономическими системами. 2012. №3. URL: http://www.uecs.ru /uecs-39-392012/item/1144—forex (дата обращения: 26.11.2014).
5. *Афанасьева В.В.* К философскому обоснованию детерминированного хаоса // URL: http://sbiblio.com/BIBLIO/archive/afanasev_k/default.aspx (дата обращения: 29.11.2014).
6. *Мусаев А.А.* Корреляционный анализ процессов изменения состояния фондовых и валютных рынков // Труды СПИИРАН. 2011. Вып. 18. С. 5–18.

7. *Kendall M.G., Stuart A.* The advanced theory of statistics. V.2. Inference and relationship // London: Ch. Griffin & Company limited. 1968. 899 p.
8. *Colby R.W.* The Encyclopedia of Technical Market Indicators. 2nd Edition // N.Y.: McGraw-Hill. 2003. 832 p.
9. *Мусаев А.А.* Мультирегрессионная оценка стоимости валютного инструмента // Известия СПбГТИ. 2015. №28(54). С. 78–85.
10. *Fogel L.J., Owens A.J., Walsh M.J.* Artificial intelligence through simulated evolution // N.Y.: John Wiley & Sons. 1966. 231 с.
11. *Демиденко Е. З.* Линейная и нелинейная регрессии // М.: Финансы и статистика, 1981. 302 с.
12. *Bolch B.W., Huang, C. J.* Multivariate statistical methods for business and economics // N.J.: Englewood Cliffs. 1974. 317 p.
13. *Мусаев А.А., Барласов И. А.* Оценивание состояния фондовых рынков на основе многомерной регрессии на скользящем окне наблюдения // Труды СПИИРАН. 2012. Вып. 19. С. 243–254.
14. *Аверченков В.И.* Эволюционное моделирование и его применение / В.И. Аверченков, П.В. Казаков. 2-е изд., стереотип // М.: ФЛИНТА. 2011. 20 0с.
15. *Курейчик В.М., Гладков Л., Курейчик В.В.* Эволюционное моделирование и генетические алгоритмы // Lambert Academic Publishing. 2011. 260 с.
16. *Карпов В.Э.* Методологические проблемы эволюционных вычислений // Искусственный интеллект и принятие решений. 2012. №4. С. 95–102.
17. *Рутковский Л.* Методы и технологии искусственного интеллекта // М.: Горячая линия–Телеком. 2010. 520 с.
18. *Mukhopadhyay A.A., Maulik U., Bandyopadhyay S., Coello C.A.* Survey of Multiobjective Evolutionary Algorithms for Data Mining: Part I // IEEE Transactions on Evolutionary Computation. 2014. vol. 18. no. 1. pp. 4–19.
19. *Mukhopadhyay A.A., Maulik U., Bandyopadhyay S., Coello C.A.* Survey of Multiobjective Evolutionary Algorithms for Data Mining: Part II // IEEE Transactions on Evolutionary Computation. 2014. vol. 18. no. 1. pp. 20–35.
20. *Мусаев А.А.* Эволюционное моделирование в задаче оптимизации управляющей стратегии // Научный вестник НГТУ. 2014. Т.56. № 3. С. 132–142.

References

1. Peters E. E. Chaos and order in the capital markets: a new view of cycles, prices, and market volatility (2nd ed.). NY: John Wiley & Sons. 1996. 288 p.
2. Williams B.M. Trading chaos. NY: John Wiley & Sons. 2002. 251 p.
3. Musaev A.A. [Modeling of quotations of trade assets]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2011. vol. 17. pp. 5–32. (In Russ.).
4. Kolodko D.V. [Not stationarity and self-similarity of the currency market Forex]. *Upravlenie jekonomicheskimi sistemami - Management of economic systems*. 2012. vol 3. Available at: http://www.uecs.ru/uecs-39-392012/item/1144--forex. (accessed 26.11.2014). (In Russ.).
5. Afanasjeva V.V. K filisofskomu obosnovaniju determinirovannogo haosa [To philosophical justification of the determined chaos]. Available at: http://sbiblio.com/BIBLIO/archive/afanasev_k /default.aspx. (accessed 29.11.2014). (In Russ.).
6. Musaev A.A. [Correlation analysis of processes of share and currency markets changes]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2011. vol. 18. pp. 5–18. (In Russ.).
7. Kendall M.G., Stuart A. The advanced theory of statistics. V.2. Inference and relationship. London: Ch. Griffin & Company limited. 1968. 899 p.
8. Colby R.W. The Encyclopedia of Technical Market Indicators. 2nd Edition. N.Y.: McGraw-Hill. 2003. 832 p.

9. Musaev A.A. [Multiregression estimation of the currency cost]. *Izvestija SPbGTI – SPbSIT News*. 2015. vol. 28(54). pp. 78–85. (In Russ.).
10. Fogel L.J., Owens A.J., Walsh M.J. Artificial intelligence through simulated evolution. N.Y.: John Wiley & Sons. 1966. 231 p.
11. Demidenko E.Z. *Linejnaja i nilinejnaja regressii* [Linear and nonlinear regressions]. Moscow: Finance and statistics. 1981. 302 p. (In Russ.).
12. Bolch B.W., Huang, C. J. Multivariate statistical methods for business and economics. N.J.: Englewood Cliffs. 1974. 317 p.
13. Musaev A.A., Barlasov I. A. [Estimation of stock markets state on the basis of multidimensional regression on the sliding watch window]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2012. vol. 19. pp. 243–254. (In Russ.).
14. Avertchenkov V.I., Kazakov P.V. *Evoljucionnoe modelirovanie i ego primenenie* [Evolutionary modeling and its application]. Moscow: FLINTA. 2011. 200 p.
15. Cureichik V.M., Gladkov L., Cureichik V.V. *Evoljucionnoe modelirovanie and geneticheskie algoritmy* [Evolutionary modeling and genetic algorithms]. Lambert Academic Publishing. 2011. 260 p. (In Russ.).
16. Carpov V.E. [Methodological problems of evolutionary calculations]. *Iskusstvennyj intellekt i prinjatie reshenij – Artificial intelligence and decision-making*. 2012. vol. 4. pp. 95–102. (In Russ.).
17. Rutkovsky L. *Metody I tehnologii iskusstvennogo intellekta* [Methods and technologies of artificial intelligence]. Moscow: Hot line–Telecom. 2010. 520 p. (In Russ.).
18. Mukhopadhyay A., Maulik U., Bandyopadhyay S., Coello C. Survey of multiobjective evolutionary algorithms for Data Mining: Part I. *IEEE Transactions on Evolutionary Computation*. 2014. vol. 18. no. 1. pp. 4–19.
19. Mukhopadhyay A., Maulik U., Bandyopadhyay S., Coello C.A. Survey of multiobjective evolutionary algorithms for Data Mining: Part II. *IEEE Transactions on Evolutionary Computation*. 2014. vol. 18. no. 1. pp. 20–35.
20. Musaev A.A. [Evolutionary modeling in a problem of the operating strategy optimization]. *Nauchnyj vestnik NGTU – Scientific bulletin NSTU*. 2014. vol. 56. no. 3. pp. 132–142. (In Russ.).

Мусаев Александр Азерович — д-р техн. наук, ведущий научный сотрудник лаборатории информационных технологий в системном анализе и моделировании, Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН), декан факультета ИТ и управления, Санкт-Петербургский государственный технологический институт (технический университет), научный консультант, ОАО Специализированная инжиниринговая компания «Севзапмонтажавтоматика». Область научных интересов: прикладная статистика, анализ данных, прогнозирование, хаотическая динамика. Число научных публикаций — 220. amusaev@technolog.edu.ru; СПИИРАН, 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ; р.т.: +7-(812)-494-9323, Факс: +7 (812)350-1113.

Musaev Alexander Azerovich — Ph.D., Dr. Sci., leading researcher, laboratory of IT in System Analysis and Modeling of St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences (SPIIRAS), dean of IT and control systems department, St. Petersburg State Technological Institute (technical university), expert, public corporation Specialized Engineering Company “Sevzapmontageautomatica”. Research interests: data analysis, complicated dynamic processes prognosis and control, stochastic chaos systems. The number of publications — 220. amusaev@technolog.edu.ru; 39, 14-th Line V.O., St. Petersburg, 199178, Russia, SPIIRAS; office phone: +7-(812)-494-9323, Fax: +7 (812)350-1113.

РЕФЕРАТ

Мусаев А.А. **Адаптивная мультирегрессионная оценка в условиях хаотических процессов валютного рынка.**

Главной особенностью динамики котировок на рынках капитала является предельно высокий уровень неопределенности, описываемой теорией хаоса. Однако важнейшей характеризацией хаотической динамики является ее свойство генерировать в себе локальные упорядоченные структуры, простейшими из которых являются локальные тренды. Искусство управления активами сводится к способности трейдера предсказать возникновение тренда и использовать его для формирования управляющей стратегии.

Важным направлением поиска упорядоченных структур в хаосе является анализ взаимных корреляционных связей, т.е. переход в область многомерного анализа данных. Наличие локальных корреляционных связей позволяет формировать скользящие оценки рыночной стоимости валютного актива по совокупности наблюдений котировок связанных с ним валютных инструментов. Текущее значение котировок может существенно отличаться от ее рыночной оценки. Это связано с наличием статистических флуктуаций, обусловленных большим числом полностью или частично неконтролируемых факторов. В этом случае рынок будет стремиться устранить данное несоответствие, что неизбежно предопределит направление движения котировки актива. Указанное свойство служит основой для построения так называемых *мультирегрессионных* осцилляторов, т.е. индикаторов состояния рынка, основанных на недооценке или переоценке текущей стоимости актива.

Очевидно, что механистический выбор регрессоров, не учитывающий реального уровня их корреляционной связи с активом, существенно снижает эффективность соответствующей управляющей стратегии. В связи с этим в настоящей работе рассмотрена задача оценки текущей стоимости актива на основе группы валютных пар, наиболее коррелированных с указанным валютным активом.

Важно заметить, что корреляционная структура хаотического многомерного процесса нестационарна и неэргодична. Однако ее изменения, в отличие от исходных процессов динамики котировок, достаточно инерционны. Это позволяет перейти к адаптивной схеме, основанной на периодическом пересчете корреляционной матрицы валютного рынка и, при необходимости, изменении состава группы регрессоров.

Оценка эффективности адаптационных технологий в задачах построения управляющих стратегий, ориентированных на функционирования в условиях хаотической динамики, является неоднозначной. Это связано с тем, что хаос, в силу нестационарности и неэргодичности, не позволяет своевременно замкнуть контур адаптивного управления. Однако в качестве вспомогательного инструмента адаптация может оказаться вполне полезной. В частности, адаптация к вариациям корреляционной структуры многомерной динамики котировок может повысить качество восстановления стоимости актива, и как следствие, поднять уровень потенциальной эффективности мультирегрессионного осциллятора.

SUMMARY

Musaev A.A. **Adaptive Multiregression Currency Estimation in the Chaotic Market Environment.**

The main feature of quotations dynamics in the capital markets is extremely high level of the uncertainty described by the theory of chaos. However the most important characterization of chaotic dynamics is its property to generate in itself the local ordered structures the simplest of which are local trends. The management skill is reduced to of the trader ability to predict a trend emergence and to use it for the operating strategy formation.

The important search direction of ordered structures in chaos is the analysis of mutual correlation relations, i.e. transition to area of the multidimensional data analysis. Existence of local correlation relations allows to form the sliding estimates of a currency values on quotations supervision set of the related currency tools. The current quotations value can significantly differ from its market assessment. It is connected with existence of the statistical fluctuations caused by a large number of unobservable factors. In this case the market will seek to eliminate this discrepancy that will inevitably predetermine the direction of the quotation movement. The specified property forms a basis for creation of so-called multiregression oscillators, i.e. the indicators of the market state based on underestimation or overestimation of the current asset cost.

It is obvious that the mechanistic choice of regressors, which isn't considering the real values of their correlated relation with an asset, significantly reduces efficiency of the corresponding operating strategy. In this regard in the real work the problem of estimation of the current asset cost on the basis of currency group which are most correlated with the specified currency is considered.

It is important to notice that correlation structure of chaotic multidimensional process is a not stationary and not ergodic. However its changes, unlike initial processes of quotations dynamics, are rather inertial. It allows to pass to the adaptive scheme based on periodic recalculation of a correlation matrix of the currency market and, if necessary, change the group of regressors.

The assessment of adaptation technologies efficiency in problems of creation of the operating strategy focused on functioning in the conditions of chaotic dynamics is ambiguous. This results from the fact that the chaos, owing to the not stationary and not ergodic, doesn't allow to close an adaptive control contour in due time. However as the auxiliary tool, adaptation can be quite useful. In particular, adaptation to variations of correlation structure of multidimensional quotations dynamics can to raise quality of asset cost restoration, and as a result, to raise the level of potential efficiency of the multiregression oscillator.

УДК 613.86;159.913

К.О. Гнидко, А.Г. Ломако, Р.Б. Жолус

ОБНАРУЖЕНИЕ ВИЗУАЛЬНЫХ КОНТАМИНАНТОВ НА ОСНОВЕ ВЫЧИСЛЕНИЯ ПЕРЦЕПТИВНОГО ХЭША

Гнидко К.О., Ломако А.Г., Жолус Р.Б. **Обнаружение визуальных контаминантов на основе вычисления перцептивного хэша.**

Аннотация. В настоящей работе предлагается подход к обнаружению широкого класса визуальных контаминантов на основе вычисления перцептивных хэшей и формирования эталонной базы данных потенциально опасных мультимедийных объектов для построения автоматической системы защиты потребителей мультимедийного контента от нежелательного воздействия на их психику и сознание.

Ключевые слова: психофизиологические воздействия, суггестия, подпороговые сообщения, скрытые изображения, распознавание образов, компьютерное зрение.

Gnidko K.O., Lomako A.G., Zholus R.B. **Detection of Visual Contaminants on the Basis of Perceptual Hash Calculation.**

Abstract. In this paper we propose an approach to the detection of a wide class of visual contaminants on the basis of visual perceptual hash calculation and formation of a reference database of potentially dangerous media objects for building an automated system to protect consumers of multimedia content from unwanted effects on their psychic and consciousness.

Keywords: psychophiziological affections, suggestion, subliminal messages, hidden images, image recognition, computer vision

1. Введение. Геополитические события последних лет, в частности череда «арабских революций», а также конфликт на Украине однозначно дают понять, что технологии воздействия на сознание окончательно вышли за пределы психологических лабораторий и стали неотъемлемой частью окружающей реальности. Сложно переоценить опасность, которую несет применение подобных технологий против Российской Федерации. Разработка эффективных мер противодействия технологиям манипулирования групповым и массовым сознанием требует всестороннего исследования особенностей психики человека, которые делают такое манипулирование возможным. Особый интерес вызывают те когнитивные искажения, которые не зависят от национальной, культурной, религиозной принадлежности и являются общими для всех представителей вида Homo sapiens.

2. Контаминация сознания как класс когнитивных искажений. Все многочисленные проявления когнитивных искажений могут быть отнесены к двум классам, первый из которых обусловлен незнанием фундаментальных закономерностей бытия или неумением применять эти знания на практике. Ко второму классу когнитивных искажений относятся случаи, когда неосознаваемые или неконтролируемые психические процессы становятся причиной нежелательной реакции индивида. В рамках настоящей работы для обозначения данного

класса ошибок мы будем использовать термин «контаминация сознания» (рисунок 1).

Рис. 1. Источники когнитивных искажений

Согласно английскому толковому словарю «Random House Dictionary» (1968 г.) глагол «to contaminate» означает «становиться грязным или непригодным для употребления в результате контакта или смешивания с чем-либо нечистым, плохим и т.д.» (с. 289). Другими словами, в случае контаминации приемлемое состояние системы становится менее приемлемым (желательным) в результате контакта системы с некоторым потенциально вредоносным агентом. Мы считаем, что многие нежелательные психические процессы и вызванные ими ошибки целесообразно рассматривать как итог воздействия на сознание и подсознание внутренних или внешних агентов-контаминантов.

Более формально определим контаминацию сознания как явление, при котором индивид формирует нежелательное суждение, испытывает нежелательные эмоции или демонстрирует нежелательное поведение вследствие неконтролируемого или бессознательного психического процесса. Под «нежелательным» мы имеем в виду тот факт, что индивид, принимающий решение, сознательно не хотел бы подвергнуться воздействию, которое в конечном счете имело место и повлияло на его решение.

Так, например, большинство преподавателей не хотели бы завышать или занижать оценки своим студентам, поддавшись влиянию своего личного субъективного отношения к ним (из-за внешнего вида, национальной принадлежности или, например, политических предпочтений). Напротив, желательной является ситуация, когда оценка является объективным отражением уровня знаний обучаемого. Однако многочисленные исследования (в частности, [1]) показали, что эффект негативного влияния сторонних факторов, не относящихся к уровню знаний обучаемого (так называемый «гало-эффект»), имеет место практически всегда. Большинство людей сознают, что реклама чаще

всего предоставляет необъективную, а порой и заведомо ложную информацию о продукте. Поэтому покупатели предпочитают, чтобы их решение о приобретении какого-либо товара не было бы инициировано просмотренной рекламой. В то же время множество примеров показывают, что в реальной жизни происходит обратное и реклама сильнейшим образом воздействует на конечное решение потенциальных потребителей [2–4].

Аналогия контаминации сознания с физическим загрязнением полезна по двум причинам. Во-первых, она подчеркивает тот факт, что сохранить «стерильность» сознания практически невозможно. При этом в реальной жизни вернуть загрязненную субстанцию в исходное «чистое» состояние как минимум непросто, а зачастую и вовсе невозможно, что, как мы полагаем, является вполне подходящей метафорой и для ментальных процессов. Контаминации сознания сложно избежать про ряду причин: неполнота знаний о законах функционирования сознания, ограниченная возможность контроля процессов, происходящих в сознании и подсознании, сложность (порой невозможность) обнаружения признаков протекания нежелательного ментального процесса.

Все проявления контаминации, в свою очередь, можно условно разделить на два типа. К первому типу относятся результаты некорректной автоматической обработки информации в сознании и подсознании индивида. Ко второму типу – последствия суперпозиции ментальных стимулов, порожденных памятью, мышлением, ощущениями, суждениями. В повседневной жизни реакции людей практически всегда определяются множеством совокупно действующих стимулов, так же, например, как оценка кандидата на вакантную должность складывается из множества факторов. Вместе с тем, большое количество проведенных исследований позволяют утверждать, что люди крайне редко способны корректно подвергнуть критическому анализу свои реакции и определить точный вклад каждого из факторов в итоговое суждение. Обычно индивид может только осознать саму мысль или ощущение, но не причины, их породившие [5, 6]. Данный эффект происходит автоматически, без контроля со стороны сознания [7], что имеет свои негативные стороны. В частности, изучение эффекта предшествования (прайминга) показало, что образы в памяти и соответствующие им модели поведения могут быть инициированы визуальной информацией, не имеющей прямого отношения к текущей ситуации.

3. Применение перцептивного хэша для автоматического обнаружения визуальных контаминантов. Мы вводим термин «ви-

зуальные контаминанты», в отношении изображений (в том числе последовательностей изображений, составляющих видеоряд), просмотр которых может приводить к нежелательному изменению мыслительных процессов и поведенческих реакций. К визуальным контаминантам могут быть отнесены, в частности, скрытые подпороговые вставки по типу «25 кадра», логотипы компаний-производителей, внедряемые в видеоряд в целях увеличения продаж, символика политических движений и т.д. В настоящей работе предлагается подход к обнаружению широкого класса визуальных контаминантов на основе вычисления перцептивных хэшей и формирования эталонной базы данных потенциально опасных мультимедийных объектов для построения автоматической системы защиты потребителей мультимедийного контента от нежелательного воздействия на их психику и сознание.

Перцептивные хэш-алгоритмы применяются для генерации на основе различных характеристик изображения индивидуальных (но не уникальных) «отпечатков» — хэшей [8]. В отличие от хэш-функций, применяемых в криптографии, перцептивные хэши можно сравнивать между собой и делать вывод о степени различия двух наборов данных (рисунок 2).

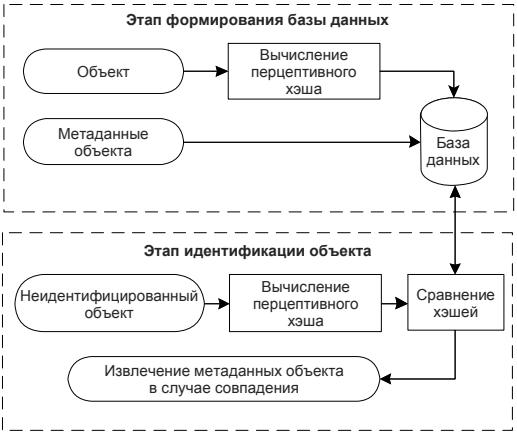


Рис. 2. Схема идентификации изображений на основе перцептивных хэшей

Перцептивные хэши устойчивы по отношению к таким преобразованиями изображений как изменение размера, изменение соотношения сторон, коррекция цветовых характеристик и незначительное вращение [9]. Данное свойство перцептивных хэшей обеспечивает обнаружение не только идентичных кадров, причем в различном разрешении, но и «похожих» с точки зрения человека образов: например,

различных картинок из одного фотосета, символов, логотипов компаний – производителей порнографической продукции. Перцептивные хэши могут успешно применяться и для обнаружения скрытых визуальных вставок в видеопотоке.

Для вычисления перцептивного хэша могут использоваться различные функции. Определим требования к ним, исходя из целевого предназначения. Введем следующие обозначения.

Пусть H — хэш-функция, которая принимает на вход объект (например, изображние) и возвращает битовую строку длины d;

x — мультимедийный объект;

$\hat{x}$ — модифицированный мультимедийный объект, перцептивно схожий с объектом x;

y — мультимедийный объект, перцептивно несхожий с x;

x' и y′ — значения перцептивных хэшей объектов x и y соответственно;

$\{0,1\}^d$ — бинарная строка длины d.

Функция перцептивного хэширования должна обладать следующими свойствами:

1. Равномерное распределение значений хэша:

$$P(H(x) = x') \approx \frac{1}{2^d}, \forall x' \in \{0,1\}^d.$$

2. Взаимная независимость для перцептивно несхожих объектов:

$$P(H(x) = x' | H(y) = y') \approx P(H(x) = x', \forall x', y' \in \{0,1\}^d.$$

3. Инвариантность относительно перцептивно схожих объектов x и $\hat{x}$:

$$P(H(x) = H(\hat{x})) \approx 1.$$

4. Высокая чувствительность к перцептивно несхожим объектам:

$$P(H(x) = H(y)) \approx 0.$$

Приведем краткое описание алгоритмов вычисления перцептивного хэша, удовлетворяющих перечисленным требованиям и наиболее часто применяемым на практике.

1. Дискретное косинусное преобразование. Пусть $x[m], m = 0, \dots, N-1$ — последовательность отсчетов сигнала длины N. Тогда дискретное косинусное преобразование:

$$X[n] = \sum_{m=0}^{N-1} c\,[n, m] \cdot x[m],$$

где матрица дискретного косинусного преобразования:

$$c[n,m] = \sqrt{\frac{2}{N}} \cdot cos\left(\frac{(2m+1)\cdot n\pi}{2N}\right), (m,n = 0,\dots,N-1).$$

Матрица $c[n,m]$ может быть вычислена заранее для любого заданного числа N, что существенно повышает скорость вычисления хэшей в случае программной реализации алгоритма.

2. Лапласиан гауссиана. Пусть $f_c(x,y)$ — функция яркости изображения в градациях серого. Тогда непрерывный лапласиан функции:

$$\nabla^2 f_c(x,y) = \nabla \cdot \nabla f_c(x,y) = \frac{\partial^2 f_c(x,y)}{\partial x^2} + \frac{\partial^2 f_c(x,y)}{\partial y^2}.$$

Фильтр Гаусса:

$$g_c(x,y) = e^{-\frac{x^2+y^2}{2\sigma^2}}.$$

Лапласиан гауссиана:

$$h_c(x,y) = \nabla^2 g_c(x,y) = \frac{x^2+y^2-2\sigma^2}{\sigma^4} \cdot e^{-\frac{x^2+y^2}{2\sigma^2}}.$$

Свертка лапласиана гауссиана с изображением:

$$[\nabla^2 g_c(x,y)] * f_c(x,y)] = \nabla^2[f_c(x,y) * g_c(x,y)].$$

3. Лучевой вектор дисперсии. Идея применения данного алгоритма состоит в построении лучевого вектора дисперсии на основе преобразования Радона, после чего к полученному вектору применяется дискретное косинусное преобразование и вычисляется хэш [10].

Пусть $I(x,y)$ — значение яркости пиксела изображения (x, y), тогда лучевой вектор дисперсии $R[\alpha]$:

$$R[\alpha] = \frac{\sum_{(x,y\in\Gamma(\alpha))} I^2(x,y)}{|\Gamma(\alpha)|} - \left(\frac{\sum_{(x,y\in\Gamma(\alpha))} I(x,y)}{|\Gamma(\alpha)|}\right)^2,$$

где угол поворота вектора проекции $\alpha = 0,1,\dots,179$; $|\Gamma(\alpha)|$ — мощность множества пикселей на линии проекции, соответствующей данному углу.

4. Вычисление хэша на основе средних значений низкочастотных характеристик изображения. Рассмотрим данный алгоритм подробнее в качестве примера. Основные шаги алгоритма описаны ниже:

Шаг 1. Удаление цветовой составляющей. Изображение конвертируется из исходного цветового пространства в градации серого.

Шаг 2. Уменьшение размера изображения для удаления высокочастотных компонент и сохранения низкочастотных. Выходной размер изображений может варьироваться (в приводимом примере используется значение 16 пикселей). Таким образом, общее число пикселей изображения после преобразования составляет $16 \times 16 = 256$ пикселей. Полученный после преобразования хэш будет соответствовать всем вариантам изображения, независимо от исходного размера и соотношения сторон (рисунок 3).

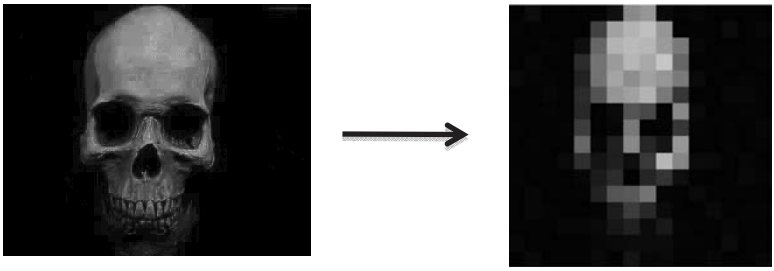


Рис. 3. Удаление высокочастотных компонент за счет уменьшения изображения

Шаг 3. Вычисление среднего значения. Для всех полученных значений градаций серого (в рассматриваемом примере - 256) рассчитывается арифметическое среднее значение.

Шаг 4. Бинаризация изображения. Для каждого пиксела изображения в градациях серого осуществляется сравнение с порогом - средним значением, вычисленным на предыдущем шаге алгоритма. Если значение пиксела превышает порог, значение пикселя заменяется на 1, в противном случае на 0 (рисунок 4).

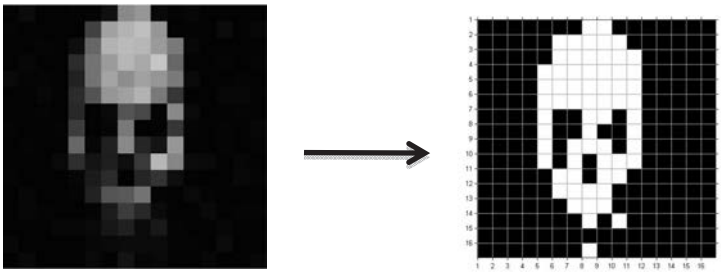


Рис. 4. Пороговая бинаризация изображения

Шаг 5. Построение хэша. Бинаризованное изображение 16×16 пикселей преобразуется в одно 256-битное значение или его 16-ричное представление. Порядок пикселей при этом не имеет значения, если он сохраняется постоянным (в применяемом алгоритме биты считываются слева направо, сверху вниз). В приведенном примере перцептивный хэш изображения черепа представлен двоичной последовательностью:

0000000110000000000001111100000000000111111000000000111111100
0000000111111100000000011111110000000001001101000000000100100
1000000000101110100000000010101110000000000110110000000000011
1100000000000001110000000000000010100000000000000000000000000
000100000000

или в шестнадцатеричном виде:

0x18007C007E00FE00FE00FE009A009200BA00AE006C0078003800140
00000100.

Для вычисления меры схожести между двумя изображениями в рассматриваемом алгоритме вычисляется расстояние Хэмминга между соответствующими хэшами:

Пусть A — алфавит конечной длины. $x = (x_1, x_2, \dots, x_n)$ и $y = (y_1, y_2, \dots, y_n)$ — строки символов одинаковой длины, где x, y ∈ A. Тогда расстояние Хэмминга Δ между x и y:

$$\Delta(x, y) = \sum_{x_i \neq y_i} 1\,, i = 1,2, \dots, n.$$

Нормированное расстояние хэмминга:

$$\Delta_{\mathrm{n}}(x, y) = \frac{1}{n} \sum_{x_i \neq y_i} 1\,, i = 1,2, \dots, n,$$

где n — длина сравниваемых строк.
Расстояние Хэмминга, таким образом, равно числу позиций, в которых векторы бинарных хэшей различны.

Продемонстрируем эффективность применения алгоритма, построенного на вычислении перцептивных хэшей, для автоматического обнаружения похожих изображений. Рассмотрим схожие между собой пары изображений, отличающиеся, тем не менее, размерами, соотношением сторон и отдельными деталями (рисунок 5).




Bear1.jpg Bear2.jpg Mouse1.jpg Mouse2.jpg

Рис. 5. Исходные изображения для сравнения на основе перцептивного хэша

Соответствующие данным изображениям значения хэша на основе средних значений среднечастотных характеристик и расстояния Хэмминга между ними приведены в таблице 1.

Таблица 1. Расстояние Хэмминга между изображениями

Наименование изображения	Значение хэша	Расстояние Хэмминга			
		Bear1	Bear2	Mouse1	Mouse2
Bear1	FFFFF8FFF2C3F013F007E3C7F387F80FF01FEF1FCF1FCF0FF20FF81FFE1FFFFF	0	27	82	81
Bear2	FFFFF0C7F043F003E10FE3C7F38FF81FF01FE01FC80FE00FF00FF80FFE1FFFFF	27	0	71	74
Mouse1	FDFFF0C7F047FDC7F8DFF93FFC1FE03FE29FE19FF13FF83FFA7FCCFFFFE7FF80	82	71	0	43
Mouse2	FDFFF8CFF80FFDCFFC3FFC7FFC3FF07FF63FF1BFF87FFC7FFE3FE47FFBDFFF3F	81	74	43	0

Установив пороговое значение для расстояния Хэмминга, можно в автоматическом режиме делать вывод о близости анализируемого изображения к заданному эталону. Так, например, установив пороговое значение 50, получим вывод о том, что из четырех представленных изображений похожи между собой Bear1 и Bear2 (27<50), а также Mouse1 и Mouse2 (43<50), что, очевидно, соответствует действительности.

4. Применение перцептивных хэшей для обнаружения скрытых подпороговых вставок в видеопотоках. В настоящее время о сведения о наличии в Российской Федерации и за рубежом действующих систем технического мониторинга каналов сети Интернет в интересах защиты от скрытой передачи вредоносной информации в открытых источниках носят единичный и отрывочный характер. Их наиболее близким аналогом является прибор ОДСВ-1 (опытный детек-

тор скрытых вставок), разработанный в 2002 году ВНИИТР совместно с Национальным исследовательским центром телевидения и радио по заказу Министерства РФ по делам печати, телерадиовещания и средств массовых коммуникаций [11]. ОДСВ-1 производит запись телевизионных программ на видеопленку, после чего осуществляет в них поиск видеовставок типа «25 кадра». Решение о вредоносности выявленных вставок выносится экспертной группой при последующем просмотре. Очевидным недостатком описанного подхода является низкая оперативность и производительность.

Известны также подходы, основанные на анализе функции интегральной яркости кадров изображения в зависимости от времени (номера кадра) и поиске локальных аномалий на графике такой функции. В качестве основного признака, позволяющего выявить наличие скрытых вредоносных вставок в видеоданных, рассматривается суммарная яркость пикселов отдельного кадра. Под скрытой визуальной вставкой понимается кадр, содержащий зрительные образы, которые должны оказывать какое-либо воздействие на человека. Для того, чтобы человек смог подвергнуться такому воздействию, кадр-вставка должна отличаться и от предыдущего, и от следующего. При этом время демонстрации скрытого кадра не должно превышать 113 мс, иначе он будет сознательно восприниматься зрителем. Для обнаружения кадра-вставки вводится вспомогательная функция, которая вычисляет отличие яркости текущего кадра от последующего по абсолютному значению. Кадры, в которых за одним скачком значения функции следует другой, помечаются как подозрительные на наличие визуальной вставки в видеопотоке.

Логическим развитием метода вычисления интегральной яркости является алгоритм обнаружения кадров-вставок на основе вычисления кадра-разности, под которым понимается результат попиксельного вычитания текущего и последующего кадров. Те области, в которых соседние кадры имеют одинаковые области изображения, при попиксельном вычитании дают в результирующем кадре черные участки. Там же, где есть отличия, появляется область, имеющая цветовую окраску. Таким образом, если два кадра представляют собой обычную последовательность фильма, то кадры-разности содержат незначительные светлые участки, которые образуются в результате движения объектов на экране. При наличии скрытой визуальной вставки результирующий кадр-разность содержит обширные цветные участки, увеличивающие суммарную дифференциальную яркость.

Принцип работы алгоритма выявления видеовставки по экстремумам суммарной дифференциальной яркости видеокадров изображен на рисунке 6.

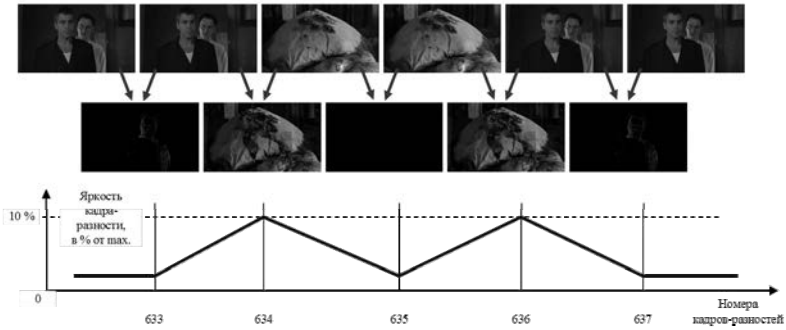


Рис. 6. Принцип работы алгоритма выявления видеовставки по экстремумам суммарной дифференциальной яркости видеокадров

Из последовательности видеокадров (верхний ряд) формируются кадры-разности (нижний ряд). Суммарная дифференциальная яркость кадров-разностей, в формировании которых были задействованы скрытые визуальные вставки, значительно превосходит соседние. Таким образом, визуальные вставки могут быть выявлены по экстремумам функции суммарной дифференциальной яркости. П-образные экстремумы характерны для визуальных вставок из одного кадра, а М-образные – для вставок двух-трех кадров.

Основным достоинством изложенного подхода является простота реализации и относительно низкая вычислительная сложность. К недостаткам следует отнести большое количество ложных срабатываний в моменты смены ракурса, высокую чувствительность метода к резким перепадам яркости (вспышка молнии, выстрел). При этом окончательное решение о наличии вредоносной вставки может быть сделано только экспертным путем.

Методы анализа кадров видеопотока, основанные на вычислении особых точек и дескрипторов изображения (SURF, SIFT и другие), устойчивы к различным искажениям и деформациям, однако существенно более ресурсоемки и, в большинстве своем, применяют защищенные патентным законодательством проприетарные алгоритмы и программные модули. Результаты проведенных нами экспериментальных исследований позволяют утверждать, что при наличии обширной базы заранее проиндексированных материалов соотношение точности

и скорости распознавания скрытых кадров-вставок в видеопотоках на основе вычисления перцептивных хэшей являются оптимальными.

В качестве примера рассмотрим фрагмент раскадровки музыкального видеоклипа, в котором присутствует скрытая полнокадровая вставка с изображением черепа (рисунок 7). Время демонстрации скрытой вставки при проигрывании клипа составляет 1/25 секунды, что не позволяет зрителю без специальной тренировки обнаружить и распознать предъявляемый таким образом эмоционально значимый символ.

кадр 136	кадр 137	кадр 138	кадр 139	кадр 140

Рис. 7. Музыкальный клип с видеовставкой

Технические характеристики анализируемого видеофайла представлены в таблице 2.

Таблица 2. Технические характеристики анализируемого видеофайла

Параметр	Значение
Формат видео:	Audio Video Interleave
Размер файла:	1,99 Мб
Длительность:	8,4 сек
Битрейт:	1 983 кБит/сек
Ширина кадра:	352 пикселов
Высота кадра:	288 пикселов
Частота демонстрации кадров:	25 кадров в секунду
Стандарт вещания:	PAL
Цветовое пространство:	YUV
Глубина цвета:	8 бит
Тип кодека:	DivX 5.1.1 (Maupiti)

Идея автоматического выявления скрытой полнокадровой вставки заключается в обнаружении единичного кадра, который перцептивно отличается как от последовательности кадров до него, так и от кадров после него. Такой кадр с высокой вероятностью является искусственно внедренным в видеопоток изображением. Для обнаружения скрытой вставки вычислим значения перцептивных хэшей всех кадров и значения расстояния Хэмминга между последовательными парами хэшей. Результат представлен в виде столбчатой диаграммы на рисунке 8. Два характерных пика на диаграмме говорят о наличии кадра (кадр №138), чей хэш «далек» от хэшей кадров, предшествую-

щих ему и следующих за ним. Данный факт позволяет строить автоматические распознаватели с настраиваемым порогом срабатывания и обнаруживать скрытые вставки в режиме времени, близком к реальному.

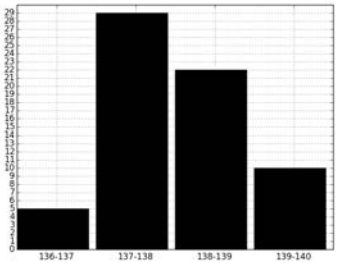


Рис. 8. Значения расстояния Хэмминга между парами хэшей кадров видеопотока, содержащего полнокадровую вставку

Важной задачей исследования является определение порогового значения расстояния Хэмминга, при котором кадр считается «отличным» от предыдущего и следующего за ним изображений. В общем случае установление порогового значения зависит от ценности защищаемого ресурса (значимости ошибок первого и второго рода), наличия сведений об априорной вероятности, появления фильтруемого контента в видеопотоке, и требований, предъявляемых к разрабатываемой системе фильтрации нежелательного видеоконтента. Сочетание данных факторов позволяет в каждом конкретном случае обоснованно применять различные статистические критерии выбора порогового значения (минимаксный критерий, критерий идеального наблюдателя и прочие).

В рамках проведенных авторами экспериментальных исследований для определения порога срабатывания программного фильтра применялся критерий Неймана-Пирсона с заданным максимальным значением вероятности ложного срабатывания $P = 0{,}1$. Исходные данные для проведения статистического эксперимента (фрагменты видеофайлов с кадрами-вставками и без них) были подготовлены с применением библиотеки компьютерного зрения OpenCV и функционально ориентированного языка программирования python. Полученные значения обработаны с помощью пакета Statistics Toolbox из состава среды инженерных вычислений Matlab. Результаты эксперимента представлены на рисунке 9.

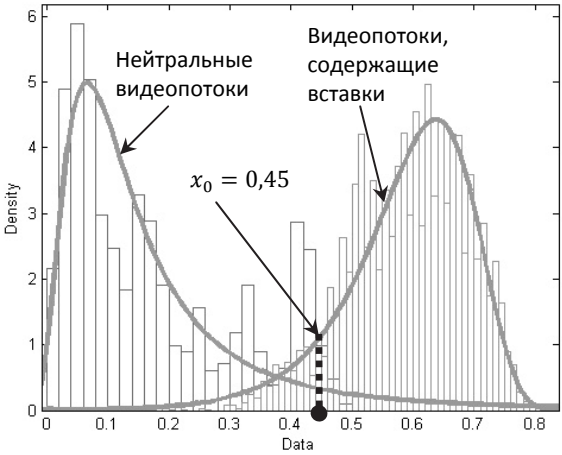


Рис. 9. Наложение гистограмм и графиков функций плотности распределения случайной величины – нормированного расстояния Хэмминга

Анализируемой случайной величиной являлось нормированное расстояние Хэмминга между соседними кадрами в видеопотоках. По оси абсцисс отложены значения рассматриваемой безразмерной случайной величины (от 0 до 1). Полученная точечная оценка позволяет установить следующий порог дифференцирования кадров:

$$x_0 = 0{,}45.$$

При этом вероятность ложных срабатываний в соответствии с критерием Неймана-Пирсона не превысит требуемой величины 0,1.

Избирательность фильтра в отношении кадров, которые не являются контаминантами, но существенно отличаются от соседних кадров видеоряда (например, смена сюжетного плана, склейка через пустой кадр), достигается добавлением в решающее правило дополнительных условий. Так, например, необходимым условием для принятия решения о наличии единичного скрытого кадра-вставки является последовательность из двух подряд идущих пиков на графике расстояния Хэмминга между парами хэшей, поскольку кадр-вставка должен существенно отличаться от соседних с ним изображений. В свою очередь, обычная смена плана приводит к появлению единичного пика на графике, что позволяет дифференцировать данные случаи. Несколько более сложной представляется задача правильного распознавания дефектов монтажа изображения. Однако и в данном случае мо-

жет быть предложено решение, основанное на сравнении хэша кадра, подозрительного на наличие визуального контаминанта, с хэшем «типового» дефекта, такого как пустой «черный» или «засвеченный» кадр.

Нейтрализация вредоносного воздействия обнаруженных скрытых кадров-вставок при условии сохранения целостности протокола передачи видеопотока может быть обеспечена путем дублирования вместо кадра-вставки предыдущего или последующего кадров из анализируемой последовательности. Однако детальное рассмотрение подходов к нейтрализации визуальных контаминантов выходит за рамки настоящей работы и является целью дальнейших исследований.

5. Заключение. Таким образом, применение перцептивного хэширования позволяет автоматизировать процесс идентификации «похожих» с точки зрения человека визуальных образов и создать базу данных хэшей потенциально опасных изображений, способных оказать нежелательное воздействие на психику и сознание. Простота реализации и относительно невысокая вычислительная сложность алгоритмов, реализующих перцептивное хэширование, делают возможным построение автоматических систем распознавания нежелательного мультимедийного контента. Вместе с тем, визуальные контаминанты являются наиболее широким, но не единственным источником суперпозиции стимулов, порождающих контаминацию сознания. Иные типы контаминантов (текстовые, аудио и т.д.) требуют разработки других подходов в рамках общей методологии распознавания образов и машинного обучения.

Противодействие контаминации второго типа, вызванной некорректной автоматической обработкой информации в сознании и подсознании индивида, с неизбежностью требует разработки моделей (в том числе динамических) индивидуального, группового и массового сознания, что является темой дальнейших исследований и публикаций.

Литература

1. *Landy D., Sigall H.* Beauty is talent: Task evaluation as a function of performer's phisical attractiveness // J. Personal. Soc. Physiol. 1974. vol. 29. pp. 299–304.
2. *Abraham M.M., Lodish L.M.* Getting the most out of advertising and promotion // Harv. Bus. Rev. 1990. vol. 68. pp. 50–58.
3. *Liebert R.M., Sprafkin J.* The early window // Elmsford. NY: Pergamon Press. 1988. Issue 3.
4. *Ryan B.J.* It works! How investment spending to adevrtising pays off // New York: 1991.
5. *Nisbett R.E., Wilson T.D.* Telling more than we can know: Verbal reports on mental processes. // Psychol. Rev. 1977. vol. 84. no. 3. 231 p.

6. *Wilson T.D. et al.* Introspection, attitude change, and attitude-behavior consistency: The disruptive effects of explaining why we feel the way we do // Orlando. FL: Academic Press. 1989. vol. 19. pp. 123-205.
7. *Bargh J.A.* Conditional automaticity: Varieties of automatic influence in social perception and cognition // Unintended thought. 1989. vol. 3. pp. 51–69.
8. *Beghdadi A. et al.* A survey of perceptual image processing methods // Signal Process. Image Commun. 2013. vol. 28. pp. 811–831.
9. *Wen Z.K. et al.* A robust and discriminative image perceptual hash algorithm // Proceedings of the 4th International Conference on Genetic and Evolutionary Computing (ICGEC 2010). 2010. pp. 709–712.
10. *Lei Y., Wang Y., Huang J.* Robust image hash in Radon transform domain for authentication // Signal Process. Image Commun. 2011. vol. 26. pp. 280–288.
11. *Паукова М.* Изобретен прибор для обнаружения 25-го кадра. URL: https://www.urfo.org/ekb/13_45277.html (дата обращения: 10.03.2015).

References

1. Landy D., Sigall H. Beauty is talent: Task evaluation as a function of performer's phisical attractiveness. *J. Personal. Soc. Physiol.* 1974. vol. 29. pp. 299–304.
2. Abraham M.M., Lodish L.M. Getting the most out of advertising and promotion. *Harv. Bus. Rev.* 1990. vol. 68. pp. 50–58.
3. Liebert R.M., Sprafkin J. The early window. Elmsford. NY: Pergamon Press. 1988. Issue 3.
4. Ryan B.J. It works! How investment spending to adevrtising pays off. New York: 1991.
5. Nisbett R.E., Wilson T.D. Telling more than we can know: Verbal reports on mental processes. *Psychol. Rev.* 1977. vol. 84. no. 3. 231 p.
6. Wilson T.D. et al. Introspection, attitude change, and attitude-behavior consistency: The disruptive effects of explaining why we feel the way we do. Orlando. FL: *Academic Press*. 1989. vol. 19. pp. 123–205.
7. Bargh J.A. Conditional automaticity: Varieties of automatic influence in social perception and cognition. *Unintended thought*. 1989. vol. 3. pp. 51–69.
8. Beghdadi A. et al. A survey of perceptual image processing methods. *Signal Process. Image Commun*. 2013. vol. 28. pp. 811–831.
9. Wen Z.K. et al. A robust and discriminative image perceptual hash algorithm. Proceedings of the 4th International Conference on Genetic and Evolutionary Computing (ICGEC 2010). 2010. pp. 709–712.
10. Lei Y., Wang Y., Huang J. Robust image hash in Radon transform domain for authentication. *Signal Process. Image Commun*. 2011. vol. 26. pp. 280–288.
11. Paukova M. Izobreten pribor dlya obnaruzheniya 25-go kadra [Device for detection of the hidden 25-th frame is invented]. Available at: https://www.urfo.org/ekb/13_45277.html (accessed: 22.01.2015). (In Russ.).

Гнидко Константин Олегович — к-т техн. наук, докторант, Военно-космическая академия имени А.Ф. Можайского. Область научных интересов: информационно-психологическая безопасность, распознавание образов, извлечение знаний из неструктурированных массивов данных. Число научных публикаций — 27. greeny598@gmail.com; ул. Ждановская, 13, 197198, Санкт-Петербург; р.т.: +7(812) 237-19-60.

Gnidko Konstantin Olegovich — Ph.D., doctoral student, Mozhaisky Military Space Academy. Research interests: information-psychological security, image recognition, data mining. The number of publications — 27. greeny598@gmail.com; 13, Zhdanovskaya street, St. Petersburg, 197198, Russia; office phone: +7(812) 237-19-60.

Ломако Александр Григорьевич — д-р техн. наук, профессор кафедры систем сбора и обработки информации, Военно-космическая академия имени А.Ф. Можайского. Область научных интересов: информационная безопасность, теоретическое и системное программирование, синтез и верификация корректности моделей программ. Число научных публикаций — 250. lomako_ag@mail.ru; ул. Ждановская 13, 197198, Санкт-Петербург; р.т.: +7(812) 237-19-60.

Lomako Aleksandr Grigor'evich — Ph.D., Dr. Sci., professor of system for collecting and processing information department, Mozhaisky Military Space Academy. Research interests: information security, theoretical and system programming, synthesis and verification of program models. The number of publications — 250. lomako_ag@mail.ru; 13, Zhdanovskaya street, St. Petersburg, 197198, Russia; office phone: +7(812) 237-19-60.

Жолус Роман Борисович — к-т биол. наук, соискатель кафедры систем сбора и обработки информации, Военно-космическая академия имени А.Ф. Можайского. Область научных интересов: информационная безопасность; моделирование социальных систем. Число научных публикаций — 10. p.glybovsky@yandex.ru; ул. Ждановская, 13, Санкт-Петербург, 197198; р.т.: +7(812) 237-19-60.

Zholus Roman Borisovich — Ph.D., applicant of system for collecting and processing information department, Mozhaisky Military Space Academy. Research interests: information security, modeling social systems. The number of publications — 10. p.glybovsky@yandex.ru; 13, Zhdanovskaya street, St. Petersburg, 197198, Russia; office phone: +7(812) 237-19-60.

РЕФЕРАТ

Гнидко К.О., Ломако А.Г., Жолус Р.Б. **Обнаружение визуальных контаминантов на основе вычисления перцептивного хэша.**

Контаминацией сознания называется ситуация, когда неосознаваемые или неконтролируемые психические процессы становятся причиной нежелательной реакции индивида. Все проявления контаминации можно условно разделить на два типа. К первому типу относятся результаты некорректной автоматической обработки информации в сознании и подсознании.

Ко второму типу – последствия суперпозиции ментальных стимулов, порожденных памятью, мышлением, ощущениями, суждениями. Контаминацию второго типа могут вызывать подпороговые вставки, логотипы компаний-производителей, эмоционально значимые символы и другие визуальные образы. В настоящей работе предлагается подход к обнаружению визуальных контаминантов на основе вычисления перцептивных хэшей и формирования эталонной базы данных потенциально опасных мультимедийных объектов. Устойчивость перцептивных хэшей к некоторым видам преобразований обеспечивает обнаружение не только идентичных кадров но и «похожих» с точки зрения человека изображений. Рассматриваются алгоритмы перцептивного хэширования, основанные на вычислении средних значений низкочастотных характеристик изображения, вычислении лапласиан гауссиана, вектора лучевой дисперсии, дискретном косинусном преобразовании. Приводится пример применения перцептивного хэширования для обнаружения скрытых подпороговых вставок в видеопотоке.

Противодействие контаминации, вызванной некорректной автоматической обработкой информации в сознании и подсознании индивида, является темой дальнейших исследований и требует разработки моделей индивидуального, группового и массового сознания.

SUMMARY

Gnidko K.O, Lomako A.G., Zholus R.B. **Detection of Visual Contaminants on the Basis of Perceptual Hash Calculation.**

Mental contamination is the situation, when unconscious or uncontrollable mental processes cause unwanted reactions of the individual. The all cases of contamination can be divided into two types. The first type is the result of incorrect automatic information processing in the conscious and subconscious. The second type – the unwanted consequences of superposition of mental stimuli generated by memory, thinking, feelings, judgments. The second type of contamination can be caused by subliminal images, logos of manufacturing companies, emotionally meaningful symbols and other visual images.

In this paper we propose an approach to the detection of contaminants on the basis of visual perceptual hash calculation and formation of a reference database of potentially dangerous multimedia objects. Stability of perceptual hashes to certain types of transforms not only makes it possible to detect identical image frames but also to find images "similar" in terms of human perception. Perceptual hashing algorithms based on the calculation of average values of low frequency features, computation of the Laplacian of the Gaussian, the radial vector dispersion and the discrete cosine transform are considered. An example of perceptual hashing usage to detect hidden subliminal images in a video stream is provided.

Contamination caused by unwanted automatic information conscious and subconscious is the subject of further research and requires the development of individual, group and mass consciousness models.

УДК 004.853

В.И. Городецкий, О.Н. Тушканова

АССОЦИАТИВНАЯ КЛАССИФИКАЦИЯ: АНАЛИТИЧЕСКИЙ ОБЗОР. ЧАСТЬ 2

Городецкий В.И., Тушканова О.Н. **Ассоциативная классификация: аналитический обзор. Часть 2.**

Аннотация. В работе продолжается рассмотрение основных результатов, моделей и методов, разработанных в области ассоциативной классификации, ориентированных на обработку данных большого объема. Дается анализ подходов, методов и алгоритмов, разработанных в области ассоциативной классификации к настоящему времени. В заключении формулируются достоинства и недостатки ассоциативной классификации как модели машинного обучения, а также дается оценка перспектив ее использования в интеллектуальном анализе больших данных.

Ключевые слова: большие данные, ассоциативное правило, ассоциативная классификация, паттерн, эмерджентный паттерн.

Gorodetsky V., Tushkanova O. **Associative Classification: Analytical Overview. Part 2.**

Abstract. The paper continues the survey of associative classification in context of big data processing. An extended overview and comparative analysis of the modern approaches, models and algorithms developed for associative classification form the main paper contents. In conclusion, the paper outlines the main advantages and drawbacks of associative classification, as well as evaluates its capabilities from big data processing perspective.

Keywords: associative classification, emerging pattern, big data.

1. Введение. Данная работа является продолжением работы [1], в которой были описаны более ранние результаты в области ассоциативной классификации. В данной работе представлены современные алгоритмы, модели и методы, разработанные в области ассоциативной классификации, которые уже в большей мере ориентированы на обработку данных большого объема.

Повторим кратко формальную постановку задачи ассоциативной классификации, приведенную более детально в первой части данной работы [1].

Пусть $\boldsymbol{D}$ – транзакционная база данных (множество данных), $D_i \in \boldsymbol{D}$ – произвольная транзакция, $\boldsymbol{X}$ – множество всех символов, которые используются для обозначения объектов (признаков, атрибутов) в транзакциях множества $\boldsymbol{D}$, $\boldsymbol{A}$ – подмножество символов из множества $\boldsymbol{X}$ и $\boldsymbol{D(A)}$ – подмножество множества транзакций из множества $\boldsymbol{D}$, каждая из которых содержит подмножество символов $\boldsymbol{A} \in \boldsymbol{X}$ в качестве подмножества. Для характеристики статистических свойств подмножества $\boldsymbol{A}$ в базе данных $\boldsymbol{D}$ используют отношение мощности n_A множества $\boldsymbol{D(A)}$ к мощности n всего множества транзакций $\boldsymbol{D}$. Эту величину

принято называть *поддержкой* (*support*) подмножества $\boldsymbol{A}$ во множестве транзакций $\boldsymbol{D}$:

$$supp(\boldsymbol{A}) = n_A / n. \quad (1)$$

Пусть даны два набора символов (объектов) $\boldsymbol{A} \in \boldsymbol{X}$ и $\boldsymbol{B} \in \boldsymbol{X}$, причем $\boldsymbol{A}$ и $\boldsymbol{B}$ не имеют общих элементов, и пусть σ и γ – вещественные числа из интервала [*0, 1*]. Говорят [2, 3], что выражение вида $\boldsymbol{A} \rightarrow \boldsymbol{B}$ есть *ассоциативное правило* с *порогом уверенности* $conf(\boldsymbol{A} \rightarrow \boldsymbol{B}) = \gamma$ и *порогом поддержки* $supp(\boldsymbol{A}) = \sigma$ (σ, γ – ассоциативное правило), если справедливы следующие неравенства:

$$n_{AB} / n \geq \sigma, \quad (2)$$

$$n_{AB} / n_A \geq \gamma, \quad (3)$$

где n_{AB} – количество транзакций во множестве $\boldsymbol{D}$, которые содержат объединение множества символов подмножеств $\boldsymbol{A}$ и $\boldsymbol{B}$. Модель ассоциативного правила, заданную условиями (2), (3), принято называть моделью типа *поддержка–уверенность*.

Подмножество (последовательность) элементов $\boldsymbol{A}$ принято называть посылкой ассоциативного правила $\boldsymbol{A} \rightarrow \boldsymbol{B}$, а подмножество (последовательность) $\boldsymbol{B}$ – его следствием. Обычно эти последовательности называют паттернами (*patterns*). В задачах ассоциативной классификации заключение правила может содержать только однолитерный паттерн, который является именем одного из классов. Поэтому в общем случае основная подзадача задачи ассоциативной классификации сводится к поиску множества (σ, γ)–ассоциативных правил для каждого класса. Эта подзадача называется обычно задачей *обучения* классификатора. Другая подзадача – это синтез классификатора на множестве найденных ассоциативных правил. Эта задача не является предметом данной работы.

В последующей части работы дается описание и сравнительный анализ основных результатов, полученных в области ассоциативного анализа к настоящему времени. В заключении формулируются достоинства и недостатки ассоциативной классификации, а также оцениваются перспективы использования методов ассоциативной классификации для интеллектуального анализа больших данных.

2. Алгоритмы ассоциативной классификации: современные подходы. Большая группа современных методов и алгоритмов генерации ассоциативных правил классификации основана на понятии эмерджентный паттерн. По сути, методы ассоциативной классифика-

ции, основанные на эмерджентных паттернах, сформировали новое направление в этой области, которое активно развивается и сегодня. Работы в этом направлении [4-7] во многом способствовали более глубокому пониманию специфики задач ассоциативной классификации.

Первой работой, в которой было введено понятие *эмерджентного паттерна* (*Emergent Pattern, EP*), далее для краткости *ЭП*, была работа [4]. В ней задача ассоциативной классификации была поставлена как задача дискриминации, т.е. как задача поиска правил, позволяющих отделить примеры одного класса от примеров другого класса. Заметим, что в работах, упомянутых ранее, прагматика ассоциативных правил как правил классификации при их генерации вообще не принималась во внимание. В работе [4], наоборот, в качестве базового критерия отбора правил ассоциативной классификации рассматривается их способность отличать примеры одного класса от примеров другого класса. В дальнейшей эволюции таких моделей эта прагматика оставалась неизменной, а совершенствование методов и реализующих их алгоритмов было направлено на повышение их вычислительной эффективности при генерации ассоциативных правил и при использовании этих правил в алгоритмах классификации.

Сформулируем понятие ЭП, следуя работе [4]. Пусть дана упорядоченная пара $\boldsymbol{D_1}$ и $\boldsymbol{D_2}$ транзакционных данных, которые либо относятся к разным классам, либо относятся к разным временным интервалам лога работы некоторой системы. Как и раньше (см. раздел 2), каждая транзакция из множеств $\boldsymbol{D_1}$ и $\boldsymbol{D_2}$ может содержать элементы (атрибуты, переменные, предметы, объекты, англ. *items*) из (линейно упорядоченного) множества (последовательности) $\boldsymbol{X}$. Рассмотрим произвольный паттерн $A \subseteq \boldsymbol{X}$, который характеризуется поддержкой σ_1 во множестве $\boldsymbol{D_1}$ и поддержкой σ_2 во множестве $\boldsymbol{D_2}$. Отношение σ_2 / σ_1 авторы работы [4] называют *коэффициентом возрастания поддержки* (*Growth rate*) паттерна $\boldsymbol{A}$ от множества данных $\boldsymbol{D_1}$ к множеству $\boldsymbol{D_2}$. Формально значение показателя *GrowthRate(A)* определяется нижеследующей формулой:

$$GrowthRate(A) = \begin{cases} 0, \text{ если } \sigma_1(\boldsymbol{A}) = 0 \text{ и } \sigma_2(\boldsymbol{A}) = 0, \\ \infty, \text{ если } \sigma_1(\boldsymbol{A}) = 0 \text{ и } \sigma_2(\boldsymbol{A}) > 0, \\ \sigma_2(\boldsymbol{A}) / \sigma_1(\boldsymbol{A}), \text{ в других случаях}. \end{cases} \quad (4)$$

Определение ЭП дается с использованием порогового значения ρ для величины *GrowthRate(**A**)*: паттерн $\boldsymbol{A}$ называется *эмерджентным паттерном* от множества $\boldsymbol{D_1}$ к множеству $\boldsymbol{D_2}$, если

GrowthRate(***A***) ≥ ρ. Таким образом, авторы вводят понятие ЭП с использованием порога ρ, чтобы выбором его значения можно было управлять их разделяющей способностью. Заметим, что понятие меры уверенности для ЭП на заданных множествах данных $\boldsymbol{D_1}$ и $\boldsymbol{D_2}$ становится ненужным, т.к. ее значение для обоих множеств равно 1, поскольку всем транзакциям каждого из множеств $\boldsymbol{D_1}$ и $\boldsymbol{D_2}$ ставится в соответствие постоянное заключение, например, метка класса или имя временного интервала.

С учетом введенных понятий и определений задача поиска ассоциативных правил в работе [4] сводится к поиску эмерджентных паттернов $\boldsymbol{A_i}$ со значением меры *GrowthRate*($\boldsymbol{A_i}$) ≥ ρ. Обратим внимание на то, что задача поиска ЭП не использует понятие поддержки, а опирается на понятие коэффициента роста поддержки. Это означает, что "хорошие" паттерны могут иметь низкое значение поддержки в обоих множествах, что приводит к значительному возрастанию вычислительной сложности задачи поиска *EP*. Причины этого обусловлены тем, что, во-первых, паттернов с низким уровнем поддержки всегда бывает очень много. Во-вторых, при поиске паттернов по условию *GrowthRate*($\boldsymbol{A_i}$) ≥ ρ принципиально нельзя воспользоваться алгоритмом *Apriori*, поскольку для *EP* не соблюдается условие монотонного уменьшения значения поддержки паттерна при увеличении его длины за счет добавления новых атрибутов.

Таким образом, данная работа показала, что поиск ассоциативных правил для задач классификации является задачей, которая, во-первых, отличается от поиска обычных ассоциативных правил, и, во-вторых, имеет не так много общего с задачей поиска правил в машинном обучении.

Для пояснения существа задачи поиска ЭП, а также для ее декомпозиции авторы используют двухмерное представление области локализации различных типов ЭП в координатах < σ_2, σ_1 >, т.е. в координатах мер поддержки паттернов во множестве данных $\boldsymbol{D_1}$ и $\boldsymbol{D_2}$. В этой области (рисунок 1) все ЭП располагаются правее прямой l_1, тангенс угла α наклона которой к оси $O\sigma_2$ равен величине $1/\rho$. Все ЭП располагаются в треугольнике *ACE*. Однако посылки правил ассоциативной классификации, должны иметь значение поддержки, превышающее минимально допустимое значение $\sigma_{2\,min}$ во множестве $\boldsymbol{D_2}$. Кроме того, во множестве $\boldsymbol{D_1}$ они должны иметь значение поддержки меньше, чем $\sigma_{2\,min}$. Паттерны с такими значениями мер поддержки σ_2

и σ_1 располагаются в прямоугольнике *BCDG*, и именно они являются целью поиска в работе [4].

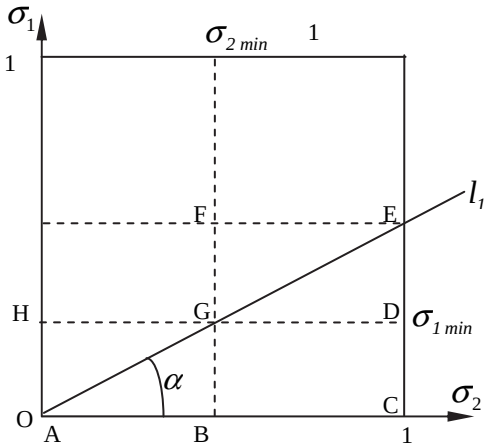


Рис. 1. Пояснение к алгоритму поиска эмерджентных паттернов

Прежде чем рассматривать алгоритм поиска таких ЭП, введем дополнительно важное понятие *интервально замкнутого множества*, использование которого позволяет авторам повысить эффективность алгоритма поиска эмерджентных правил. Заметим, что хотя авторы претендуют на авторство во введении этого понятия, оно давно и хорошо известно в алгебре, в частности, в теории алгебраических структур. Однако использование этого понятия для эффективного поиска ассоциативных правил, несомненно, принадлежит авторам [4].

Интервально замкнутое множество $\Im$ определяется следующим образом. Пусть дана упорядоченная пара множеств подмножеств [***L, R***], $\boldsymbol{L} = \{\boldsymbol{A}_i\}_{i=1}^{k}$ и $\boldsymbol{R} = \{\boldsymbol{B}_j\}_{j=1}^{r}$, при этом любое подмножество $\boldsymbol{A_i} \subseteq \boldsymbol{B_j}$ для некоторого значения индекса *j*, а все подмножества $\boldsymbol{A} \in \{\boldsymbol{A}_i\}_{i=1}^{k}$, как и все подмножества $\boldsymbol{B} \in \{\boldsymbol{B}_j\}_{j=1}^{r}$, являются несравнимыми по отношению включения. Тогда множество *всех* подмножеств $\{\boldsymbol{Z}_l\}_{l=1}^{s}$, такое, что для любого $\boldsymbol{Z}_l$ найдется пара подмножеств $\boldsymbol{A_i} \in \boldsymbol{L}$ и $\boldsymbol{B_i} \in \boldsymbol{R}$, для которых выполнено условие $\boldsymbol{A_i} \subseteq \boldsymbol{Z_l} \subseteq \boldsymbol{B_j}$, называется интервально замкнутым множеством подмножеств $\Im$=[**L**, **R**] с грани-

цами **L** и **R**. Множества подмножеств $\boldsymbol{L}=\{\boldsymbol{A}_i\}|_{i=1}^{k}$ и $\boldsymbol{R}=\{\boldsymbol{B}_j\}|_{j=1}^{r}$ называются *правой* и *левой границами* множества $\Im$ соответственно. Заметим, что границы множества $\Im$ принадлежат ему. Для заданных границ интервально замкнутое множество определяется единственным образом. Справедливо и обратное: если множество $\Im$ является интервально замкнутым, то его нижние и верхние границы определяются единственным образом.

Приведем пример интервально замкнутого множества подмножеств (рисунок 2) [4]. Пусть заданы границы множества: левая (нижняя) **L**={{θ}}, где θ – символ пустого множества, и правая (верхняя) **R** = {{a_1,a_2,a_3}, {a_1,a_4}}. Тогда интервально замкнутое множество $\Im$=[{{θ}}, {{a_1,a_2,a_3}, {a_1,a_4}}]={{a_1}, {a_2}, {a_3}, {a_1,a_2}, {a_1,a_3}, {a_2,a_3}, {a_1,a_2,a_3}, {a_4}, {a_1,a_4}}.

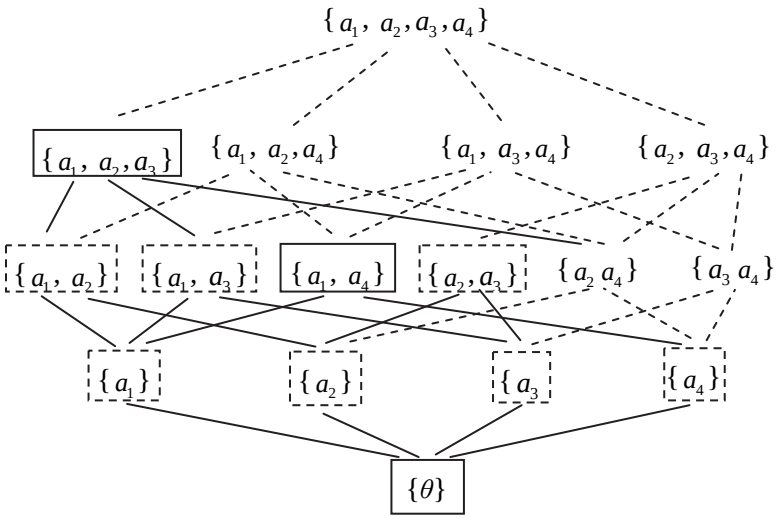


Рис. 2. Пример интервально замкнутого множества. Его элементы обведены прямоугольниками. Элементы нижней и верхней границ обведены прямоугольниками со сплошными сторонами

Способ построения интервально замкнутого множества по заданным его границам можно пояснить с помощью диаграммы Хассе частично упорядоченного множества $\Im$=[**L**, **R**], рассматриваемого в примере. Максимальные и минимальные элементы входят в интер-

вально замкнутое множество $\Im$. Остальные его элементы формируются как множества подмножеств его максимальных элементов, каждое из которых содержит в себе хотя бы один минимальный элемент. В рассматриваемом примере **L**={{θ}}, **R** = {{a_1, a_2, a_3}, {a_1, a_4}}. П содержит всего 4 разных элемента, а именно, a_1, a_2, a_3 и a_4. Диаграмма Хассе множества всех подмножеств, которые могут быть образованы из этих элементов, представлена на рисунке 3. В ней всего 16 элементов вместе с пустым множеством. Элементы нижней и верхней границ рассматриваемого примера на этой диаграмме обведены прямоугольниками со сплошными границами. Остальные элементы множества $\Im$=[**L**, **R**] на диаграмме Хассе, представленной на рисунке 3, обведены прямоугольниками с пунктирными границами. Отношения включения на множестве элементов $\Im$ показаны сплошными линиями. Таким образом, 9 множеств, входящих в интервально замкнутое множество $\Im$ (не считая пустого множества), задаются всего тремя граничными множествами.

Понятие интервально замкнутого множества паттернов (подмножеств) для поиска эмерджентных ассоциативных правил оказывается весьма полезным при построении эффективных алгоритмов их поиска, поскольку

а) множество всех паттернов, удовлетворяющих заданному ограничению на минимальное значение поддержки, является интервально замкнутым, и

б) алгоритм их поиска оказывается возможным свести к манипуляциям только с элементами границ интервально замкнутого множества. Например, вместо рассмотрения 9 паттернов в приведенном примере можно будет ограничиться рассмотрением только двух максимальных паттернов (в данном примере нижняя граница формируется пустым множеством, которое не соответствует какому-либо паттерну).

Среди интервально замкнутых семейств множеств авторы выделяют три специальных случая. Множество $\Im$=[**L,R**] называется *левокорневым*, если его левая граница состоит из одноэлементного множества. Если его правая граница состоит их одноэлементного множества, то оно называется *правокорневым*. В общем случае, если одноэлементными являются или множества, задающие правую границу, или множества, задающие левую границу, или оба эти множества являются корневыми, то множество $\Im$=[**L,R**] называют *корневым*. Авторы ограничиваются рассмотрением именно корневых интервально замкнутых множеств паттернов.

Опишем теперь кратко алгоритм поиска ЭП, предложенный в работе [4]. Он опирается на утверждение о том, что *множество всех паттернов (подмножеств элементов), которые имеют поддержку не меньше, чем некоторый заданный порог* $\sigma_{2\,min}$, *является левокорневым интервально замкнутым* множеством [4]. Аналогичное утверждение справедливо также для множества паттернов, которые имеют поддержку меньше, чем некоторый заданный порог $\sigma_{2\,min}$. Такое множество является *правокорневым* интервально замкнутым множеством. Если обратиться к рисунку 2, то правокорневым будет множество всех ЭП, для которых пары [σ_2, σ_1] отвечают точкам треугольника *ABG*. Множество всех ЭП для множества транзакций $\boldsymbol{D_2}$, которое отвечает четырехугольнику *BCEG*, будет левокорневым интервально замкнутым множеством паттернов. Напомним, что все эти паттерны являются эмерджентными от множества данных $\boldsymbol{D_1}$ к множеству данных $\boldsymbol{D_2}$ (поскольку эта область лежит ниже прямой l_1).

Обозначим *правую* границу множества паттернов для данных $\boldsymbol{D} \in \{ \boldsymbol{D}_1, \boldsymbol{D}_2 \}$ с поддержкой больше порога σ_{min} символом $LB_{\sigma_{min}}(\boldsymbol{D})$ (*Large Border*), а левую границу множества паттернов для данных $\boldsymbol{D}$ с поддержкой меньше порога σ_{min} – символом $SB_{\sigma_{min}}(\boldsymbol{D})$ (*Small Border*). Множество всех паттернов первого типа обозначим символом $\Im = \left(LB_{\sigma_{min}}, \boldsymbol{D}\right) = \left[\{\theta\}, LB_{\sigma_{min}}(\boldsymbol{D})\right]$, а множество всех паттернов второго типа обозначим символом $\Im = \left(SB_{\sigma_{min}}, \boldsymbol{D}\right) = \left[SB_{\sigma_{min}}(\boldsymbol{D}), \{I\}\right]$, где $\{\theta\}$ – пустое множество, а $\{I\}$–универсальное множество, т.е. множество всех паттернов, которые могут быть составлены из символов множества *I*.

Эмерджентные паттерны, которые являются кандидатами на использование их в качестве посылок ассоциативных правил, разделяющих множества $\boldsymbol{D_1}$ и $\boldsymbol{D_2}$, должны быть σ_{2min}– паттернами множества данных $\boldsymbol{D_2}$. Но они не должны быть одновременно σ_{1min}– паттернами множества данных $\boldsymbol{D_1}$. Суть алгоритма, предложенного в рассматриваемой работе, состоит в поиске таких и только таких ЭП. Эффективность алгоритма обеспечивается тем, что он не оперирует с самими множествами $\Im = \left(LB_{\sigma_{1min}}, \boldsymbol{D_1}\right)$ и $\Im = \left(LB_{\sigma_{2min}}, \boldsymbol{D_2}\right)$. Он оперирует только их верхними и нижними границами.

Дадим теперь описание алгоритма поиска ЭП. Он строится на основе трех базовых алгоритмов, называемых далее алгоритмами 1, 2 и

3 соответственно. Рассмотрим их кратко. Детальные их описания и обоснования могут быть найдены в работе [4], где эти алгоритмы представлены в псевдокоде.

Алгоритм 1. В основу этого алгоритма положен алгоритм *Max-Miner*, предложенный в работе [8]. Он позволяет для заданного множества данных $\boldsymbol{D} \in \{\boldsymbol{D_1}, \boldsymbol{D_2}\}$ эффективно строить паттерны максимальной длины со значением поддержки не менее, чем заданное значение порога. Если говорить в терминах решаемой задачи поиска ЭП, то он позволяет быстро находить верхние (правые) границы $LB_{\sigma_{2min}}(\boldsymbol{D_2})$ и $LB_{\sigma_{1min}}(\boldsymbol{D_1})$ множеств паттернов для множеств данных $\boldsymbol{D_2}$ и $\boldsymbol{D_2}$ соответственно. Напомним, что левой границей для обоих множеств данных является одноэлементное множество $\{\theta\}$. Как отмечается в работе [8], алгоритм *Max-Miner* может эффективно отыскивать паттерны максимальной длины, содержащие до 13 элементов. Он строится на основе стандартного алгоритма *Apriori* [9]. Дополнительно к нему, алгоритм *Max-Miner* использует процедуру *просмотра вперед* (*look ahead*), генерируя в опережающей манере паттерны большей длины, чем те, которые обычно генерируются алгоритмом *Apriori* на его текущих шагах. Увеличение эффективности достигается за счет того, что просмотр вперед позволяет отсечь неперспективные направления продолжения поиска максимальных паттернов на более ранних шагах.

Итак, *алгоритм 1* отыскивает множества максимальных паттернов $LB_{\sigma_{2min}}(\boldsymbol{D_2})$ и $LB_{\sigma_{1min}}(\boldsymbol{D_1})$ при заданных значениях нижних порогов поддержки σ_{2min} и σ_{1min} соответственно. В общем случае они будут иметь вид:

$$LB_{\sigma_{1min}}(\boldsymbol{D_1}) = [\{\theta\}, \{\boldsymbol{A_1}, \boldsymbol{A_2}, \ldots, \boldsymbol{A_s}\}] \quad (5)$$

$$LB_{\sigma_{2min}}(\boldsymbol{D_2}) = [\{\theta\}, \{\boldsymbol{B_1}, \boldsymbol{B_2}, \ldots, \boldsymbol{B_k}\}]. \quad (6)$$

Как указывалось ранее, дальнейший поиск множества ЭП сводится к поиску границ множества, которое включает в себя паттерны множества $\Im = (LB_{\sigma_{2min}}, \boldsymbol{D_2})$, из которого удалены паттерны, входящие во множество $\Im = (LB_{\sigma_{1min}}, \boldsymbol{D_1})$. Этот поиск реализуется с использованием алгоритмов 2 и 3.

Алгоритм 2. Он реализует стандартную операцию поиска множества минимальных элементов (другими словами, левой границы) теоретико–множественной разности двух интервально замкнутых левокорневых множеств паттернов, имеющих в качестве левой границы

пустое множество { θ }. Дополнительно к этому, первый элемент искомой разности должен быть одновременно и правокорневым множеством паттернов. Заметим, что в алгоритме поиска ЭП он применяется для поиска левой границы разности двух интервально замкнутых множеств вида $[\{\theta\},\{\boldsymbol{B}_j\}]$ и $[\{\theta\},\{\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_i\}]$, где $\boldsymbol{B}_j \in \{\boldsymbol{B}_1, \boldsymbol{B}_2, \ldots, \boldsymbol{B}_k\}$ (см. (5) и (6)).

Авторы рассматривают два варианта реализации этой операции. В первом варианте сначала генерируются оба множества паттернов по их границам $[\{\theta\},\{\boldsymbol{B}_j\}]$ и $[\{\theta\},\{\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_k\}]$, а затем из первого множества паттернов удаляются те паттерны, которые одновременно содержатся и во втором множестве. Далее из полученного множества удаляются элементы, которые не являются минимальными в смысле частичного порядка по включению множеств. Второй вариант алгоритма, более эффективный, чем первый, отличается от него тем, что в нем построение верхней границы выполняется рекурсивно по отношению к множествам $\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_k$ с удалением элементов, не являющихся минимальными, на каждом шаге. Таким способом достигается снижение мощности общего множества паттернов, генерируемых и просматриваемых в процессе построения границ разности двух левокорневых интервально замкнутых множеств указанного вида.

Алгоритм 3. Этот алгоритм использует алгоритмы 1 и 2. Он находит все ЭП, которые отвечают прямоугольнику BCDG, представленному на рисунке 1. Дадим краткое описание алгоритма 3. Его псевдокод можно найти в [4].

Входом алгоритма являются два интервально замкнутых множества паттернов (5) и (6) с паттернами множеств $LB_{\sigma_{2min}}(\boldsymbol{D}_2)$ и $LB_{\sigma_{1min}}(\boldsymbol{D}_1)$ в качестве правых границ множеств паттернов для данных $\boldsymbol{D}_2$, и $\boldsymbol{D}_1$ соответственно. Они находятся с помощью алгоритма 1, описанного выше.

Алгоритм реализуется последовательно для значений индексов $j \in 1,2,\ldots,k$ нижеследующим образом.

1. Для паттерна $\boldsymbol{B}_j$ строятся паттерны $\hat{\boldsymbol{A}}_1 = \boldsymbol{B}_j \cap \boldsymbol{A}_1$, $\hat{\boldsymbol{A}}_2 = \boldsymbol{B}_j \cap \boldsymbol{A}_2, \ldots, \hat{\boldsymbol{A}}_k = \boldsymbol{B}_j \cap \boldsymbol{B}_k$.

2. С помощью алгоритма 1 строится множество максимальных паттернов для найденного множества паттернов $\{\hat{\boldsymbol{A}}_1, \hat{\boldsymbol{A}}_2, \ldots, \hat{\boldsymbol{A}}_k\}$. Обозначим полученное множество символом $SB_{\sigma_{2min}}(\hat{\boldsymbol{A}}\boldsymbol{B}_j)$.

3. С помощью алгоритма 2 находится левая граница разности двух интервально замкнутых левокорневых множеств паттернов $\left\langle\{\theta\},\{\boldsymbol{B}_j\}\right\rangle$ и $\left\langle\{\theta\},\{SB_{\sigma_{min}}(\hat{\boldsymbol{A}}\boldsymbol{B}_j)\}\right\rangle$ и полученный результат, а именно множество минимальных паттернов (множество паттернов левой границы) этой разности добавляется в искомое множество ЭП.

Напомним, что описанный алгоритм имеет целью поиск ЭП, ля которых значения мер поддержки отвечают прямоугольнику *BCDG* (см. рисунок 2). Авторы замечают, что поиск ЭП, отвечающих треугольнику *ABG*, является вычислительно сложной задачей ввиду того, что в этой области паттерны, получаемые на основе данных множества $\boldsymbol{D_1}$, обладают низким значение поддержки, а потому их может быть катастрофически много. Наоборот, паттерны, которые получаются для множества данных $\boldsymbol{D_1}$, будут отвечать треугольнику *GDE*, обычно бывает немного и их можно проверить простым перебором.

Работа [4] рассматривает только вопрос о том, как находятся ЭП с заданными областями значений поддержки во множествах $\boldsymbol{D_2}$ и $\boldsymbol{D_1}$. Но эта работа не рассматривает, каким образом полученные ЭП используются далее в алгоритме ассоциативной классификации. Этому вопросу посвящена работа [5]. Дадим краткое описание технологии построения ассоциативного классификатора, представленной в этой работе.

Отметим, что одна из особенностей задачи классификации с использованием ЭП в качестве посылок правил состоит в том, что приходится использовать большое число правил, каждое из которых может покрывать только небольшое число примеров обучающей выборки. То же самое имеет место, как правило, и в режиме работы с новыми данными. Это означает, что с помощью таких правил нельзя строить алгоритмы классификации типа голосования правил. В отличие от этого, правила классификации, генерируемые большинством других методов с большим значением покрытия, можно рассматривать как самостоятельные классификаторы. Если каждый из них имеет вероятность правильной классификации больше, чем 0,5, то в таком случае в соответствии с теоремой Кондорсе результат голосования сходится с вероятностью 1 к правильному решению при увеличении числа правил [10, 11]. Поэтому для таких правил схемы голосования работают обычно хорошо. В отличие от этого, каждый ЭП работает правильно на очень небольшой доле обучающих данных, а вероятность правильной классификации с помощью ЭП на всем множестве данных может быть менее 0,01. Поэтому было бы правильнее интерпретировать ЭП как некоторые более удобные новые признаки, каждый из которых все еще не может рассматриваться как "хороший" классификатор. Именно поэтому авторы работы [5] рассматривают задачу построения классификаторов на основе ЭП как самостоятельную задачу.

Полагая, что алгоритм поиска ЭП имеется (см. [4]), авторы работы [5] рассматривают процедуру построения ассоциативного классификатора, в которой поиск ЭП является одной из готовых процедур. Такой подход позволяет хорошо понять достоинства и недостатки ассоциативной классификации на основе ЭП. Опишем кратко предложенную процедуру построения ассоциативного классификатора.

Первым ее шагом является преобразование обучающих данных к булевой форме, когда каждый элемент транзакции $\boldsymbol{A} \subseteq \boldsymbol{X}$ представляется булевой переменной, которая принимает значение *true*, если соответствующий элемент множества (последовательности) $\boldsymbol{X}$ в транзакции $\boldsymbol{A}$ присутствует, и значение *false*, в противном случае. Для перехода к такому представлению данных авторы используют самый простой метод – разбиение на интервалы числовых атрибутов с введением новой пропозициональной переменной для каждого интервала. Заметим, что такой вариант преобразования исходных типов данных задачи к булевой форме может привести к заметному увеличению размерности итогового пространства атрибутов, в котором придется далее решать задачу обучения классификатора.

На втором шаге отыскиваются ЭП для каждого класса, когда все множество данных обучения разбивается на два подмножества. Одно из них – это подмножество данных класса, для которого отыскиваются ЭП, а второе – это данные всех остальных классов. Заметим, что такой подход является типичным для задач классификации, в которой число классов больше двух.

Третий шаг построения классификатора является ключевым, и именно он содержит основные особенности. Его целью является агрегирование ЭП, имеющее целью преобразование подмножеств ЭП каждого класса в более выразительные структуры для увеличения их *дискриминационных* возможностей.

Сначала авторы рассматривают индивидуальные разделительные возможности ЭП. Эти возможности, обычно, очень ограничены. Например, если ЭП покрывает 3% примеров данных и при этом он классифицирует их правильно с вероятностью, например, 0,8, то вероятность правильного предсказания этим ЭП на всем множестве примеров обучающих данных будет порядка 0,024 [5]. Для оценки индивидуальных разделительных возможностей ЭП вводят формулу:

$$Score(E_i, C, s) = \frac{Growth\_Rate(E_i)}{Growth\_Rate(E_i)+1} \times support_C(E_i), \quad (7)$$

где $\boldsymbol{E_i}$ – ЭП, разделительные возможности которого оцениваются функцией (7) для примера s по отношению к классу $\boldsymbol{C}$, а величина $support_C(E_i)$ есть поддержка паттерна $\boldsymbol{E_i}$ в классе $\boldsymbol{C}$.

Авторы обращают внимание на то, что в выражении $\{Growth\_Rate(E_i)/ (Growth\_Rate(E_i)+1)\} \times support_C(E_i)$ первый сомножитель примерно равен *условной вероятности* события "*пример s, в котором имеется ЭП* $\boldsymbol{E_i}$*, принадлежит классу C*", а второй сомножитель – это доля примеров класса *C*, которые содержат данный ЭП. Сумму всех таких величин по множеству всех ЭП класса *C* авторы предлагают рассматривать в качестве величины, характеризующей разделительную силу построенного множества ЭП для этого класса:

$$Score\ (C, s) = \sum_{E_i \subseteq s, E_i \in E(C)} \frac{Growth\ \_\ Rate\ (E_i)}{Growth\ \_\ Rate\ (E_i) + 1} \times support\ _C\ (E_i) \qquad (8)$$

Этот выбор авторов очень уязвим. Эта величина была бы равна полной вероятности класса *C* в том случае, если бы ЭП класса на выборке класса *C* были бы независимыми случайными величинами. А это, в свою очередь, может иметь место в том и только в том случае, когда каждый паттерн покрывает некоторое множество данных, которые не покрываются ни одним другим паттерном. Такой случай нереален, поэтому мотивация авторов в этой части является неубедительной. Авторы понимают, что оценка разделительной способности множества паттернов в виде (8) не является вполне корректной. В реальных ситуациях сумма в (8) будет напрямую зависеть от числа ЭП, сгенерированных для класса, а также от меры зависимости паттернов класса между собой. Поэтому строить выбор классификатор по максимуму этой величины нельзя. Для ослабления названных недостатков меры (8) при ее использовании в качестве атрибута классификации авторы предлагают нормировать эту величину по множеству всех классов. В вычислении нормированных значений величин (8) для всех классов состоит основная задача третьего шага построения алгоритма ассоциативной классификации, рассматриваемого здесь.

Эта нормировка выбирается следующим образом. Нормирующий коэффициент строится таким образом, чтобы ослабить влияние того факта, что разные классы могут иметь разное количество ЭП. В качестве нормирующего коэффициента используется величина, которую авторы называют *base_score*(*C*). Она вычисляется как *медиана* множества значений величин *Score*(***C***,*s*) для всех тренировочных данных класса *C*. Медиана отвечает значению величины *Score*(***C***,*s*) для того примера данных класса *C*, для которого 50% всех тренировочных данных этого класса имеют значения больше него. Это значение *Score*(***C***,*s*) для конкретного примера берется в качестве значения нормирующего коэффициента *base_score*(***C***) класса ***C*** при вычислении

нормированного значения функции типа (8), используемой в алгоритме классификации:

$$Norm\_Score(\boldsymbol{C},s) = score(\boldsymbol{C},s,) / base\_score(\boldsymbol{C}). \quad (9)$$

Авторы, однако, замечают, что такой выбор не является обязательным. Их эксперименты показали, что выбор в качестве медианы любого примера в пределах (50–85)% всех примеров не сказывается существенно на качестве классификатора.

Что касается самого алгоритма CAEP, то в нем выбор величины *base_score*(*C*) выполняется автоматически с помощью последовательного повышения величины нижнего порога ρ для значения меры *Growth_Rate*($\underline{E_i}$) и выбора конкретного примера в качестве медианы.

Данный алгоритм проверен авторами на большом числе реальных данных. По их утверждению он показал хорошие свойства как по эффективности поиска ассоциативных правил классификации, так и по качеству решения самих задач классификации данных. В частности, экспериментальные результаты авторов показали, что алгоритм CAEP обладает лучшими свойствами по точности классификации по сравнению с классическим методом C4.5, а также по сравнению с методом CBA, который был рассмотрен ранее.

Однако авторы работ [4, 5], хотя и выражают оптимизм по поводу вычислительной эффективности ассоциативной классификации на основе ЭП, понимают, что достигнутого уровня вычислительной эффективности алгоритмов синтеза таких классификаторов явно недостаточно для работы с *большими данными*. Их исследования в течение последующего десятилетия позволили им существенно улучшить как эффективность, так и точность модели обучения ассоциативной классификации, в основе которой лежит понятие ЭП. Это достигнуто как за счет расширения понятия эмерджентного паттерна, так и за счет повышения эффективности алгоритмов их поиска.

В работе [7] авторы алгоритма CAEP вводят понятие *скачкообразного эмерджентного паттерна* (*Jumping Emergent Pattern*, *JEP*), далее для краткости *СЭП*, который отличается от понятия ЭП в следующем: скачкообразный эмерджентный паттерн – это ЭП, который в одном множестве данных имеет нулевое значение поддержки, а в другом – строго положительное.

Вообще говоря, правила классификации, которые имеют посылку с ненулевой поддержкой только в одном классе, рассматривались в алгоритмах AQ [12, 13], GK2 [14] и в других алгоритмах, которые брали за основу модусы сходства и различия Д.С.Милля [15, 16]. Новизна идеи использования понятия СЭП состоит только в том, что он рас-

сматривается в контексте задачи ассоциативной классификации. Что касается алгоритма поиска СЭП и построения алгоритма ассоциативной классификации по найденному множеству СЭП (авторы обозначают этот алгоритм аббревиатурой *JEP*), то отличия здесь не носят принципиального характера. Как и в случае алгоритма CAEP, алгоритм *JEP* использует понятие правой и левой границ множества ЭП и сводит поиск СЭП к поиску паттернов, задающих левую границу множества всех ЭП. Такой выбор обосновывается тем, что (а) паттерны, задаваемые элементами левой границы, имеют наименьшую длину на множестве других сравнимых с ними ЭП, а значит, и наибольшее значение меры поддержки по сравнению с ними, и (б) любое подмножество паттернов левой границы уже может не являться СЭП.

Алгоритм поиска СЭП в алгоритме *JEP* строится авторами во многом по аналогии с алгоритмом CAEP и с частичным его использованием. В нем, как уже говорилось, поиск СЭП сводится к поиску элементов левой границы всех ЭП. Эти СЭП авторы называют *наиболее выразительными* (*most expressive*) паттернами. Множество таких паттернов находится для каждого класса $\boldsymbol{C_p} \in \{\boldsymbol{C_1}, \boldsymbol{C_2}, \ldots, \boldsymbol{C_q}\}$, для которого в качестве альтернативного класса выступает множество остальных классов $\overline{\boldsymbol{C}}_{\boldsymbol{p}} \in \{\boldsymbol{C_1}, \boldsymbol{C_2}, \ldots, \boldsymbol{C_q}\} \setminus \boldsymbol{C_p}$. Соответственно, в качестве обучающих данных для класса C_p выступает множество D_p, а для класса $\overline{C}_p$ обучающие данные формируются как объединение обучающих данных всех классов, кроме данных класса C_p. Обозначим это множество данных символом $\overline{D}_p$. С учетом введенных обозначений для каждого класса пары $\{C_p, \overline{C}_p\}$ строится множество наиболее выразительных паттернов *MEJEP*($D_p, \overline{D}_p$). На этом обучение ассоциативного классификатора в модели СЭП заканчивается.

Рассмотрим, каким образом в алгоритме *JEP* работает классификатор. Обозначим, как и ранее, символом s новую транзакцию, для которой нужно предсказать класс принадлежности. Для решения этой задачи авторы работы [7] для каждого класса $\boldsymbol{C_p} \in \{\boldsymbol{C_1}, \boldsymbol{C_2}, \ldots, \boldsymbol{C_q}\}$ предлагают вычислять значение меры, которую они называют *коллективным влиянием СЭП* (*collective impact, CI*). Эта мера вычисляется по такой формуле:

$$CI(C_p) = \sum_{i:\ СЭП_i \in ME-JEP(C_p, \overline{C}_p)\ \&\ СЭП \subseteq s} supp_{D_p}(СЭП_i) \cdot \qquad (10)$$

Решение принимается в пользу того класса C_p, для которого значение коллективного влияния $CI(C_p)$ максимально. Заметим, что такая модель принятия решения имеет много общего со схемой взвешенного голосования.

Основные различия алгоритмов *JEP* и *CAEP* состоит в том, что в них по-разному выбираются множества ЭП, а также по-иному строятся алгоритмы классификации. В алгоритме *JEP* не используется понятие *GrowthRate*. Оба алгоритма имеют примерно одинаковую точность принятия решений. Оба они, по заключению авторов, превосходят по точности и по вычислительной эффективности алгоритм *CBA* и классический алгоритм *C4.5*.

Алгоритмы *CAEP* и *JEP* имеют свои достоинства и недостатки. Например, алгоритм *CAEP* использует паттерны с некоторым пороговым значением поддержки (например, 1%), а СЭП с такой поддержкой могут не существовать. Хотя алгоритм *JEP* создан как некое развитие алгоритма *CAEP*, он не дает принципиального улучшения как вычислительной эффективности, так и по точности ассоциативной классификации по сравнению с аналогичными характеристиками алгоритма *CAEP*, но авторы выражают определенный оптимизм по поводу его возможностей.

В работе [17] используется некоторая модификация СЭП – СЭП с подсчетом встречаемости (англ. *Jumping Emerging Patterns with Occurrence Counts).* Для нахождения СЭП в работе также используется алгоритм с построением границ. Авторы экспериментально показывают, что такое расширение СЭП хорошо работает в области классификации изображений.

В своих более поздних работах (см., например, [6] и др.) авторы алгоритмов *CAEP* и *JEP* признают, что хотя число наиболее выразительных СЭП, которые генерируются алгоритмом *JEP*, намного меньше, чем общее их число, тем не менее, даже для данных небольшого объема и размерности их получается слишком много. Например, в одной из задач для данных общим объемом 1000 примеров, которые описываются двадцатью атрибутами, общее число СЭП оказалось равным 32244. Из них было отобрано 2754 наиболее выразительных паттернов, однако и это число слишком велико для эффективной и устойчивой классификации. Попытка дальнейшего снижения числа используемых СЭП может потребовать продолжения фазы обучения на этапе построения классификатора. Для преодоления этих недостатков авторы пошли по пути поиска ЭП других типов, которые присутствуют в данных в меньшем числе, но обладают большей дискриминационной силой.

Авторы [18] также отказываются от использования СЭП в пользу обычных ЭП и экспериментально показывают, что ЭП дают более точную классификацию.

Первый, достаточно естественный, шаг – это введение ограничений на минимальное значение поддержки СЭП, что обеспечивает некоторый минимальный уровень покрытия им тренировочных данных. Такой паттерн авторы называют строгим СЭП и формально определяют его следующим образом [6]. Паттерн $\boldsymbol{A} \in \boldsymbol{X}$ называется *строгим скачкообразным ЭП* (*Strong Jumping Emergent Pattern, SJEP*) из множества $\boldsymbol{D_1}$ во множество $\boldsymbol{D_2}$, далее ССЭП, если для него выполнены следующие два условия:

1. $supp_{D_1}(\boldsymbol{A}) = 0$ и $supp_{D_2}(\boldsymbol{A}) \geq \delta$;

2. Любое собственное подмножество элементов паттерна $\boldsymbol{A}$ не удовлетворяет условию 1.

Этот шаг позволил авторам снизить число и повысить выразительность паттернов, на базе которых строится ассоциативный классификатор. Вторым, не менее важным шагом в направлении повышения вычислительной эффективности является введение ими для представления множества паттернов специальной структуры, которая обеспечивает не только эффективное их хранение, но также эффективный просмотр и поиск. Эта структура названа авторами *деревом контрастных паттернов* (*Contrast Pattern Tree Structure, CPT structure*) [6]. Следует заметить, что идея использования такого дерева заимствована ими из работы [19], где похожая структура была предложена для *представления возрастающих паттернов* (*FP-growth Tree*). В структуре *FP-growth Tree* множество паттернов задается в префиксной форме, в которой все паттерны, имеющие общий префикс, представляются общим путем из корня дерева до узла, которому соответствует последний общий символ паттернов (последний символ общего префикса). В работе [6] паттерны генерируются и структурируются точно так же, как и в алгоритме *FP-growth*. Однако содержание каждого узла структуры *CPT* намного богаче, чем содержание узла в дереве *FP-growth Tree*. Оно имеет целью обеспечить лучшую поддержку процессов формирования множеств ССЭП и представление информации о них.

Еще одно новшество структуры *CPT* состоит в том, что представляемые в нем данные предварительно упорядочиваются по отношению $\prec$ следующим образом. Пусть имеются два множества данных, и $\boldsymbol{D_2}$, относящиеся к разным классам, и множество атрибутов этих данных $\boldsymbol{X} = \{x_1, x_2, \ldots, x_n\}$, где любое $x_i \in \{x_1, x_2, \ldots, x_n\}$ есть одно-

элементное множество, символ паттерна, объект и т.п. Пусть, в соответствии с определением ССЭП, δ– это порог для минимального значения поддержки ССЭП. Определим величину $SupportRatio(x_i)$, называемую отношением поддержек элемента x_i в двух множествах $\boldsymbol{D_1}$ и $\boldsymbol{D_2}$, следующим образом [6]:

$$SupportRatio(x_i) = \begin{cases} 0, \ \textit{если} \ [supp_{D_1}(\{x_i\}) < \delta] \wedge [supp_{D_2}(\{x_i\}) < \delta], \\ \infty, \ \textit{если} \ [supp_{D_1}(\{x_i\}) = 0] \wedge [supp_{D_2}(\{x_i\}) \geq \delta] \vee \\ \quad [supp_{D_1}(\{x_i\}) \geq \delta] \wedge [supp_{D_2}(\{x_i\}) = 0], \\ \max\left(\dfrac{supp_{D_2}(\{x_i\})}{supp_{D_1}(\{x_i\})}, \ \dfrac{supp_{D_1}(\{x_i\})}{upp_{D_2}(\{x_i\})}\right), \ \textit{иначе}. \end{cases} \quad (11)$$

Очевидно, что чем больше значение функции *SupportRatio(x_i)*, тем лучше разделяющие свойства элемента $\{x_i\}$. Обычно эта функция принимает значения больше 1, кроме тех случаев, когда оба значения поддержки меньше значения порога δ, но такие паттерны являются бесполезными. Если же значение функции для одноэлементного паттерна $\{x_i\}$ равно ∞, то такой одноэлементный паттерн уже является ССЭП.

Отношение $\prec$ определяется в работе [6] следующим образом: для пары одноэлементных паттернов $\{x_i\}$ и $\{x_j\}$ справедливо $\{x_i\} \prec \{x_j\}$, если *SupportRatio(x_i)*> *SupportRatio(x_j)* или *SupportRatio(x_i)= SupportRatio(x_j), но* $x_i < x_j$ лексикографически.

Далее предполагается, что все элементы паттерна упорядочены в нем по введенному порядку, так что любой паттерн является упорядоченным списком. Аналогичным образом вводится порядок на множестве паттернов произвольной длины. Говорят, что паттерн $\{x_1, x_2, \ldots, x_m\} \prec \{y_1, y_2, \ldots, y_n\}$, если или (1) существует номер i, $1 \leq i \leq m$, такой, что когда $1 < j < i$ $x_j = y_j$, но $x_j \prec y_i$, или (2) для любого j из интервала $1 < j < m$ $x_j = y_j$, но $m < n$.

Дерево контрастных паттернов СРТ строится таким образом, чтобы в нем все ветви, исходящие из корня, представляли паттерны, упорядоченные *слева направо* (чем левее паттерн, тем он важнее) и *от корня дерева вниз* – в каждой ветви (более важные одноэлементные паттерны находятся в ветви дерева ближе к корню). Каждый узел дерева представляет собой упорядоченное слева направо подмножество одноэлементных паттернов, каждому из которых поставлено в соответствие значение функции поддержки во множествах $\boldsymbol{D_1}$ и $\boldsymbol{D_2}$. Рассмотрим при-

мер дерева *CPT*, заимствованный из работы [6], который поясняет введенное понятие порядка на множестве одноэлементных паттернов, а также структуру *CPT*. Пусть имеются данные, представленные в таблице 1 [6], в которой в последнем столбце экземпляры данных записаны в порядке, предусмотренном введенным отношением $\prec$. Этот порядок легко строится на основе данных таблицы.

Таблица 1. Обучающие данные. Пример заимствован из работы [6]

ID	Метка класса	Данные	Данные, упорядоченные отношением $\prec$
1	$\boldsymbol{D}_1$	{*a*,*c*,*d*,*e*}	{*e*,*a*,*c*,*d*}
2	$\boldsymbol{D}_1$	{*a*}	{*a*}
3	$\boldsymbol{D}_1$	{*b*,*e*}	{*e*,*b*}
4	$\boldsymbol{D}_1$	{*b*,*c*,*d*,*e*}	{*e*,*b*,*c*,*d*}
5	$\boldsymbol{D}_1$	{*a*,*b*}	{*a*,b}
6	$\boldsymbol{D}_1$	{*c*,*e*}	{*e*,c}
7	$\boldsymbol{D}_1$	{*a*,*b*,*c*,*d*}	{*a*,*b*,*c*,*d*}
8	$\boldsymbol{D}_1$	{*d*,*e*}	{*e*,*d*}

Структура данных *CPT* для этого множества имеет вид, представленный на рисунке 3а. В этом дереве все примеры с общим префиксом имеют общий отрезок пути из корня до некоторого узла. Например, экземпляры данных {*e*,*a*,*c*,*d*}, {*e*,*b*}, {*e*,*b*,*c*,*d*}, {*e*,*c*} и {*e*,*d*} имеют общий префикс {*e*}, поэтому все они имеют в дереве общий участок пути, который проходит через узел *e*. Экземпляры данных {*e*,*b*}, {*e*,*b*,*c*,*d*} имеют общий префикс {*e*,*b*}, поэтому общим участком в дереве для них является отрезок пути от корня до узла *b*. Каждому узлу $\{x_i\}$ дерева поставлены в соответствие значения функции $supp_{\boldsymbol{D}_i}(\{x_i\})$ во множествах $\boldsymbol{D_1}$ и $\boldsymbol{D_2}$ для паттернов, отвечающих множеству узлов на пути от корня структуры до соответствующего узла включительно. Заметим, что значения этой функции представлены в терминах абсолютных значений, т.е. в числе экземпляров данных, в котором соответствующий паттерн встречается в них. Каждому узлу поставлено в соответствие два таких числа. Одно из них представляет значение функции $supp_{\boldsymbol{D}_1}(\{x_i\})$, а другое – значение функции $supp_{\boldsymbol{D}_2}(\{x_i\})$. В работе [6] дано строгое описание алгоритма построения *CPT* для данных двух множеств $\boldsymbol{D_1}$ и $\boldsymbol{D_2}$ в псевдокоде, поэтому здесь описание этого алгоритма опускается.

Структура *CPT* для представления данных используется далее алгоритмом генерации ССЭП. Формальное описание этого алгоритма

дано в работе [6] в псевдокоде, поэтому здесь ограничимся только его содержательными пояснениями на примере.

Для идентификации некоторого поддерева *CPT* в интересах последующего поиска и просмотра его содержания будем использовать имя корневого узла этого поддерева в форме префикса, т.е. в форме последовательности имен узлов от корня *CPT* до корневого узла поддерева. Например (рисунок 3а), идентификатор *R.e* ссылается на поддерево, корнем которого является узел *e*. Легко понять, например, о каком поддереве идет речь, если его корень имеет идентификатор *R.ea* или *R.ab*. Значение поддержки любого паттерн, заданного своим префиксом, во множествах $\boldsymbol{D_1}$ и $\boldsymbol{D_2}$, указывается в соответствующем узле *CPT* явно (см. рисунок 3а) для обоих множеств.

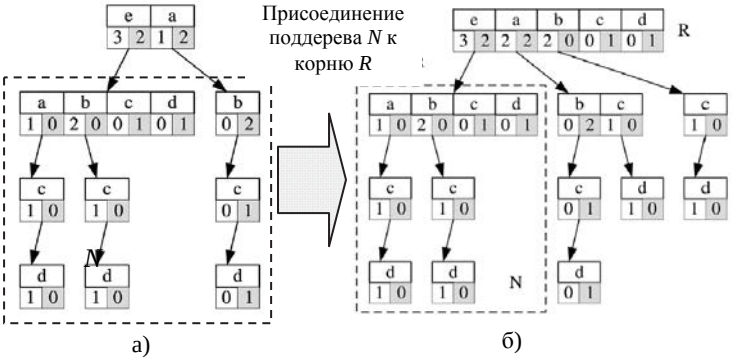


Рис. 3. а) Исходное представление данных табл. 1 в виде дерева контрастных паттернов. Рисунок заимствован из работы [6]; б) Скорректированное представление данных табл. 1 в виде дерева контрастных паттернов. Рисунок заимствован из работы [6]

В качестве входных данных алгоритм генерации множества ССЭП использует дерево *CPT*. Заметим, что в *CPT*–дереве на рисунок 3а одинаковые паттерны встречаются в нескольких узлах, но с разными префиксами. Например, это касается паттерна $\{a\}$, который встречается в узлах *R.a* (с пустым префиксом) и *R.ea* (с префиксом *e*). В таком дереве перечислять паттерны неудобно. Поэтому для решения задачи генерации паттернов авторы предлагают выполнять преобразование *CPT* –дерева, представляющего данные множеств $\boldsymbol{D_1}$ и $\boldsymbol{D_2}$, в более удобную форму. Заметим, что в алгоритме генерации ССЭП это преобразование является частью алгоритма и вызывается им по мере необходимости.

Поясним только общую идею этого преобразования, поскольку детальное его объяснение заняло бы слишком много места. В рабо-

те [6] этот алгоритм представлен в псевдокоде. Преобразование выполняется обходом узлов *CPT* – структуры в соответствии со стратегией *поиск в глубину*. Это означает, что обход узлов (на рисунок 3а) выполняется путем последовательного просмотра паттернов $\{e\}$, $\{e,a\}$, $\{e,a,c\}$, $\{e,a,c,d\}$, $\{e,b\}$, $\{e,b,c\}$, $\{e,b,c,d\}$, $\{e,c\}$ $\{e,d\}$, $\{a\}$, $\{a,b\}$, $\{a,b,c\}$ и $\{a,b,c,d\}$. Но заметим, что поддерево с корнем *R.e* содержит паттерны, которые встречаются и в поддереве *R.a*. Очевидно, что поддержка таких паттернов должна быть суммирована по множеству всех узлов, где они встречаются. С этой целью авторами работы [6] используется операция *присоединения поддеревьев* (*merge of subtrees*). Например, на рисунок 3а поддерево, обозначенное символом *N* (в ранее принятых терминах это дерево обозначалось бы символом *R.e*) присоединено к корню дерева *R* на рис. 3б. В полученной структуре в корне *R* суммируются значения поддержек одноэлементных паттернов, которые они имеют в исходном и в присоединенном деревьях. В рассматриваемом примере оказывается достаточным однократное использование операции присоединения поддеревьев. В алгоритме генерация ССЭП эта операция используется рекурсивно, когда это необходимо, по мере обхода узлов *CPT*–дерева в глубину и слева направо.

Напомним, что последовательность представления одноэлементных паттернов в любом узле дерева, как и порядок их следования вдоль ветвей дерева, определяются отношением $\prec$, а этот порядок на множестве одноэлементных паттернов вычисляется на основе данных множеств $\boldsymbol{D_1}$ и $\boldsymbol{D_2}$. Использование такого упорядочивания одноэлементных паттернов в узлах и в глубину дерева приводит к тому, что самые выразительные (самые полезные) паттерны располагаются ближе к корню *CPT*–дерева и в его ветвях, лежащих левее. С учетом этого свойства в алгоритме генерации ССЭП обычно требуется просмотр только небольшой части этого дерева в левой верхней его области. Покажем это на примере.

Примем в примере минимально допустимое значение этой поддержки $\delta = 2$. Поиск ССЭП начинается с анализа паттерна $\{e\}$, отвечающего корню *CPT*–дерева *R.e*, который имеет поддержку [3,2] во множествах $\boldsymbol{D_1}$ и $\boldsymbol{D_2}$ соответственно. Для него функция $SupportRatio(e)$ равна *1,5*, и поскольку она не равна бесконечности, то паттерн $\{e\}$ не может быть ССЭП–паттерном. Но так как функция $SupportRatio(e)$ не является антимонотонной по длине паттерна, то возможно, что символ *e* встречается в ССЭП большей длины. Поэтому просмотр ветвей поддерева *R.e* в глубину следует продолжить. При таком просмотре далее нужно анализировать узел *R. ea*. Но паттерн $\{ea\}$ имеет максимальное значение поддержки в одном из множеств, равное 1, а потому не является ССЭП. То же самое будет иметь место и для всех паттернов, которые могут находиться в поддереве с корнем

R. ea, поскольку значение поддержки вдоль ветвей дерева не может увеличиваться. Поэтому поддерево *R. ea* далее не рассматривается.

Очередным узлом для анализа является узел *R. eb* [2,0]. Ему соответствует паттерн {*eb*}, который удовлетворяет определению ССЭП для множества $\boldsymbol{D_1}$. Таким образом, один ССЭП уже найден, это паттерн {*eb*}[2,0]. Дальнейший просмотр ветвей поддерева *R. eb* в глубину не имеет смысла, поскольку паттерны, которые в нем могут встретиться, имеют большую длину, чем паттерн {*eb*}, а потому не могут иметь большее значение поддержки.

Следующий узел для анализа – это узел *R.ec*. С учетом всех случаев паттернов, которые начинаются символом *c*, узел *R. ec* имеет поддержку [2,1] во множествах $\boldsymbol{D_1}$ и $\boldsymbol{D_2}$, соответственно. Паттерн {*ec*} не является ССЭП, поскольку его поддержка не равна нулю в одном из множеств $\boldsymbol{D_1}$ или $\boldsymbol{D_2}$. Но это не исключает, что более длинный паттерн по одной из ветвей, исходящих из узла *R. ec*, может таковыми оказаться для множества $\boldsymbol{D_1}$, в котором он имеет допустимое значение поддержки. Действительно, паттерн {*ecd*}[2,0] удовлетворяет определению ССЭП в пользу множества $\boldsymbol{D_1}$.

При дальнейшем обходе дерева будет найден еще один ССЭП, а именно {*ab*}[0,2]. Таким образом, для множества данных, представленных в табл. 1, алгоритм находит следующие ССЭП: {*eb*}[2,0], {*ecd*}[2,0] и {*ab*}[0,2].

Псевдокод алгоритма генерации ССЭП по данным, представленным структурой CPT–дерева, может быть найден в работе [6].

Достоинство описанного алгоритма состоит в его эффективности. Кроме того, его достоинство состоит также в том, что он выполняет поиск ССЭП одновременно для двух множеств $\boldsymbol{D_1}$ и $\boldsymbol{D_2}$ за один проход всей базы данных. В этом его неоспоримое преимущество по сравнению с алгоритмами поиска ЭП и СЭП, в которых этот поиск делается поочередно с использованием представления множества эмерджентных паттернов с помощью верхней и нижней границ.

Дальнейшие усилия авторов данного направления были направлены на поиск других типов эмерджентных паттернов, которые могли бы лучше подойти для построения ассоциативных классификаторов и которые могут быть эффективно сгенерированы. Один из предложенных вариантов несколько обобщает понятие ССЭП, допуская, что поддержка таких паттернов не обязательно должна быть равна нулю в одном из классов. Оправданием такого допущения является возможность зашумления ССЭП, которое приведет к незначительному отклонению поддержки ССЭП в одном из классов от нуля. Это предположение представляется достаточно разумным, поскольку на тестовых данных практически всегда ССЭП имеет ненулевую поддержку в соответствующем классе. Заметим, что то же самое имеет место при тестировании правил, полученных алгоритмами индуктивного обучения, например, алгорит-

мами *AQ* [12, 13] или *GK2* [14]. Такие паттерны авторы называют *эмерджентными паттернами, устойчивыми к шуму* (*Noise-tolerant EPs, NEPs*). В формальном определении NEP допускается, что порог его поддержки в одном из множеств не превышает заданной малой величины, а порог поддержки в другом, наоборот, не меньше, чем заданное пороговое значение. Авторы вводят также понятие *обобщенного эмерджентного паттерна, устойчивого к шуму* (*Generalized Noise–tolerant EP, GNEP*). Обобщение состоит в том, что вместо традиционной меры различительной силы паттернов, задаваемой функцией *GrowthRate*, авторы допускают использование некоторых функций от значений продержки в двух множествах. Однако конструктивность и полезность такого понятия авторами не мотивируется.

Что касается использования полученных ССЭП в алгоритме ассоциативной классификации, то в этом отношении авторы не предлагают чего-либо нового и рассматривают варианты, аналогичные тем, что были предложены ими и другими исследователями в данной области.

Эффективность и качество алгоритмов генерации ССЭП, а также их использование в задачах классификации были тщательно исследованы авторами на большом количестве наборов данных. Полученные экспериментальные результаты сравнивались с результатами, полученными для наиболее известных алгоритмов обучения и классификации, например, для *CBA*, *C4.5* и его версиями, улучшенными за счет бустинга, и другими алгоритмами. Во всех случаях алгоритм *SJEP* оказывался существенно лучше по эффективности и показывал, в среднем, лучшие результаты по точности классификации. Причем эти оценки были получены на десятках различных наборов данных из UCI–репозитория [20]. Алгоритм *SJEP* сравнивался также с алгоритмом *JEP* и показал в среднем десятикратное ускорение решения задач и сравнимую точность при меньшем числе используемых паттернов. Несколько более осторожно авторы делают заключение о перспективности использования паттернов, которые являются обобщением ССЭП, а именно эмерджентных паттернов, устойчивых к шуму и их обобщений.

Очевидно, что модели ассоциативной классификации на основе различных видов эмерджентных паттернов являются важным шагом в области построения эффективных моделей классификации при работе с большими данными. Представляется, однако, что на текущий момент они исчерпали свои возможности по дальнейшему повышению эффективности. Одной из причин для такого заключения относительно модели, основанной на использовании ЭП, является необходимость сведения в ней любых данных к модели булевых данных. Такое преобразование всегда приводит к заметному увеличению размерности задачи, а значит, ставит дополнительные ограничения на возможности такого подхода при работе с большими данными.

3. Заключение. Появление модели эмерджентного паттерна, заимствованного из классической теории индуктивного обучения, было существенным шагом вперед в области поиска ассоциативных правил классификации. Это обусловлено тем, что в этой модели сама задача поиска ассоциативных правил классификации была сформулирована с учетом прагматики задачи классификации, суть которой состоит в том, что требуется построить правила, которые могли бы отделить экземпляры данных одного класса от экземпляров данных других классов. Эта прагматика была явно встроена в понятие ЭП. Особенно точно это выражено в понятии ССЭП. Это позволило авторам в дальнейшем сосредоточиться на эффективности алгоритмов поиска ЭП. Другое большое достижение в этой части – это введение структуры *CPT*–дерева для экономного представления данных обучения и построения эффективных алгоритмов их использования в процессах генерации ССЭП, отвечающих посылкам выразительных правил ассоциативной классификации. Обратим внимание на тот факт, что структура *CPT*–дерева, хотя первоначальная идея ее использования принадлежит и не авторам работ в области алгоритмов генерации ССЭП, является чрезвычайно продуктивной идеей в области обучения классификации. По-видимому, это дерево при некоторой его модификации может быть использовано в алгоритмах поиска минимальных правил в задачах индуктивного обучения при решении задач типа оптимального покрытия [12-14]. Оно может найти и другие применения.

Однако вопрос о пределах и перспективах использования моделей и алгоритмов ассоциативной классификации в задачах анализа больших данных в настоящее время вряд ли имеет однозначный ответ. К основному недостатку такой модели следует отнести ее ограниченные возможности при работе с гетерогенными данными сложной структуры, с данными, представленными текстами на естественном языке, изображениями и т.п. Модель ассоциативной классификации хорошо подходит для работы с дискретными данными (булевыми, номинальными и целочисленными). В случае других типов данных принципиальной проблемой становится проблема трансформации реальных данных к дискретной модели. Пути решения этой проблемы известны, однако их вряд ли можно считать удовлетворительными. В любом случае известные способы такого преобразования приводят к существенному увеличению размерности данных, возможно в десятки и сотни раз. А проблема размерности является ключевой проблемой больших данных и без их дискретизации.

Другая проблема, которая осложняет использование имеющихся моделей и алгоритмов ассоциативной классификации, вызвана тем фактом, что связь, которая называется ассоциацией, носит чисто синтаксический характер. Она является ненаправленной связью и потому ее нельзя интерпретировать как причинно–следственную связь, если

не обосновывать это методами или не использовать метрики, которые специально для этого разрабатываются. Во многих случаях ассоциация возникает как следствие "третьих факторов", от которых зависят и посылка и заключение ассоциативного фактора. Источником семантически нелепых связей, которые при этом могут быть обнаружены, является чисто синтаксический характер ассоциаций.

Анализируя способы построения классификаторов на основе ассоциативных правил, полученных с помощью процедур обучения, легко заметить, что такие способы базируются на здравом смысле, на эвристиках, на введении специальных метрик для оценки разделительной способности классификаторов. Для каждого варианта выбора модели объединения решений, даваемых различными правилами (по крайней мере, из тех вариантов, которые были рассмотрены в данной работе), всегда легко построить пример, когда предложенный вариант совсем не подходит, или когда предложенная эвристика не работает. Эта проблема достаточно хорошо известна, и она детально анализируется в теории объединения решений, которые вырабатываются множеством классификаторов, а каждое ассоциативное правило может интерпретироваться как простейший пример классификатора. Она называется проблемой разнообразия классификаторов (см., например, ее анализ в работе [21]). От успешности решения этой проблемы во многом будет зависеть успешность модели ассоциативной классификации при работе с большими данными.

Еще одна проблема, которая свойственна задаче принятия решений, в частности, задаче классификации при работе с большими данными, это учет контекста обучающих данных и, соответственно, контекста экземпляра данных, подлежащего классификации. Эта проблема исследуется особо в современной литературе по интеллектуальной обработке больших данных, в частности, она исследуется в задачах классификации, решаемых рекомендующими системами (см., например, работу [22]). Как известно, одним из предлагаемых решений проблемы учета контекста в задачах классификации больших данных является использование онтологий. Представляется, что модель ассоциативной классификации может успешно интегрироваться с моделью онтологического представления гетерогенных данных большого объема и размерности.

В заключение можно утверждать, что интеграция идей индуктивного обучения и ассоциативного анализа данных для построения моделей принятия решений при работе с большими данными, в частности, ассоциативной классификации представляется достаточно перспективной идеей. Эта интеграция, возможно, уже в ближайшее время сможет дать новый толчок в направлении эффективного решения проблемы обучения в задачах классификации и синтеза классификаторов применительно к большим данным. Однако успех будет сильно зави-

сеть от того, насколько эффективно удастся преодолеть отмеченные выше проблемы.

Литература

1. *Городецкий В.И., Тушканова О.Н.* Ассоциативная классификация: аналитический обзор. Часть 1 // Труды СПИИРАН. 2015. №1(38). С. 183–203.
2. *Городецкий В.И., Самойлов В.В.* Ассоциативный и причинный анализ и ассоциативные байесовские сети // Труды СПИИРАН. 2009. №9. С. 13–65.
3. *Adamo J.-M.* Data Mining for Association Rules and Sequential Patterns // Springer. 2000.
4. *Dong G., Li J.* Efficient Mining of Emerging Patterns: Discovering Trends and Differences // Proc. of the KDD'99. 1999. pp. 43–52.
5. *Dong G., Zhang X., Wong L., Li J.* CAEP: Classification by Aggregating Emerging Patterns // Proc. of the DS'99. 1999. pp. 30–42.
6. *Fan H., Ramamohanarao K.* Fast Discovery and the Generalization of Strong Jumping Emerging Patterns for Building Compact and Accurate Classifiers // IEEE Trans. Knowl. Data Eng. 2006. vol. 18(6). pp. 721–737.
7. *Li J., Dong G., Ramamohanarao K.* Making use of the most expressive jumping emerging patterns for classification // Proc. of the Fourth Pacific-Asia Conference on Knowledge Discovery and Data Mining, Kyoto, Japan. 2000. pp. 220–232.
8. *Bayardo Jr. R.J.* Efficiently Mining Long Patterns from Databases // Proc. of the SIGMOD Conference. 1998. pp. 85–93.
9. *Agrawal R., Sricant R.* Fast Algorithm for Mining Association rules // Proc. of the 20th Intern. Conference on Very Large Databases. Santiago, Chile. 1994. pp. 68–77.
10. *Condorcet N.C.* Essaisur l'application de l'analyse a la probabilite des decisions rendues a la pluralité des voix. Paris: Imprimerie Royale. 1785.
11. Condorcet's jury theorem. Wikipedia.org: the free encyclopedia // URL: http://en.wikipedia.org/wiki/Condorcet's_jury_theorem (дата обращения 20.06.2014 г.).
12. *Michalski R.S.* On the Quasi-Minimal Solution of the General Covering Problem// Proc. of the V International Symposium on Information Processing (FCIP-69), Bled, Yugoslavia. 1969. vol. A3. pp .125–128.
13. *Michalski R.S.* A Theory and Methodology of Inductive Learning. Machine Learning, vol.1 // Eds. Carbone J.G., Michalski R.S., Mitchel T.M. Tigoda. Palo Alto. 1983. pp. 83–134.
14. *Gorodetsky V., Karsaev O., Samoilov V.* Direct Mining of Rules from Data with Missing Values // Studies in Computational Intelligence. Chapter in book. Eds. Lin T.Y., Ohsuga S., Liau C.J., Hu X.T., Tsumoto S.. Foundation of Data Mining and Knowledge Discovery. Springer. 2005. vol. 6. pp. 233–264.
15. *Милль Дж.Ст.* Система логики силлогистической и индуктивной: Изложение принципов доказательства в связи с методами научного исследования Пер. с англ. // Изд. 5, испр. и доп. М.: ЛЕНАНД, 2011.
16. Пять канонов Джона Милля. Vikent.ru – портал И.Л. Викентьева // URL: http://vikent.ru/enc/834/ (дата обращения 20.06 2014 г.).
17. *Kobylinski L., Walczak K.* Efficient Mining of Jumping Emerging Patterns with Occurrence Counts for Classification // Transactions on Rough Sets XIII. LNCS 6499. 2011. pp. 73–88.
18. *Sherhod R., Judson P.N., et al.* Emerging Pattern Mining To Aid Toxicological Knowledge Discovery // Journal of Chemical Information Modeling. 2014. no. 54 (7). pp 1864–1879.
19. *Han J., Pei J., Yin Y.* Mining frequent patterns without candidate generation // Proc. of the ACM SIGMOD Intern. Conf. on Management of Data. 2000. pp. 1–12.

20. *Blake C.L., Murphy P.M.* UCI Repository of machine learning database. University of California, Department of Information and Computer Science. Irvine, CA. 1998 // URL: http://www.cs.uci.edu/mleam/mlrepository.html (дата обращения 20.06.2014).
21. *Городецкий В.И., Серебряков С.В.* Методы и алгоритмы коллективного распознавания // Автоматика и Телемеханика. 2008. № 11. С. 3–40.
22. *Gorodetsky V., Samoylov V., Serebryakov S.* Ontology–based Context–dependent Personalization Technology // Proc. of the WI/IAT 2010. Toronto. 2010. pp. 278–283.

References

1. Gorodetsky V.I., Tushkanova O.N. [Associative classification: analytical overview. Part 1]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2015. no. 1(38). pp. 183–203. (In Russ.).
2. Gorodetsky V.I., Samoylov V.V. [Associative and causal analysis and associative Bayesian networks]. *Trudy SPIIRAN - SPIIRAS Proceedings*. 2009. no. 9. pp. 13–65. (In Russ.).
3. Adamo J.-M. Data Mining for Association Rules and Sequential Patterns. Springer, 2000.
4. Dong G., Li J. Efficient Mining of Emerging Patterns: Discovering Trends and Differences. Proc. of the KDD'99. 1999. pp. 43–52.
5. Dong G., Zhang X., Wong L., Li J. CAEP: Classification by Aggregating Emerging Patterns. Proc. of the DS'99. 1999. pp. 30–42.
6. Fan H., Ramamohanarao K. Fast Discovery and the Generalization of Strong Jumping Emerging Patterns for Building Compact and Accurate Classifiers. IEEE Trans. Knowl. Data Eng. 2006. vol. 18(6). pp. 721–737.
7. Li J., Dong G., Ramamohanarao K. Making use of the most expressive jumping emerging patterns for classification. Proc. of the Fourth Pacific-Asia Conference on Knowledge Discovery and Data Mining, Kyoto, Japan, 2000. pp. 220-232.
8. Bayardo R.J.Jr. Efficiently Mining Long Patterns from Databases. Proc. of the SIGMOD Conference. 1998. pp. 85–93.
9. Agrawal R., Sricant R. Fast Algorithm for Mining Association rules. Proc. of the 20th Intern. Conference on Very Large Databases. Santiago, Chile. 1994. pp. 68–77.
10. Condorcet N.C. Essaisur l'application de l'analyse a la probabilite des decisions rendues a la pluralité des voix. Paris: Imprimerie Royale. 1785.
11. Condorcet's jury theorem. Wikipedia.org: the free encyclopedia. Available at: http://en.wikipedia.org/wiki/Condorcet's_jury_theorem (accesses: 20.06.2014).
12. Michalski R.S. On the Quasi-Minimal Solution of the General Covering Problem. Proc. of the V International Symposium on Information Processing (FCIP-69), Bled, Yugoslavia. 1969. vol. A3. pp .125–128.
13. Michalski R.S. A Theory and Methodology of Inductive Learning. Machine Learning, vol.1. Eds. Carbone J.G., Michalski R.S., Mitchel T.M. Tigoda. Palo Alto. 1983. pp. 83–134.
14. Gorodetsky V., Karsaev O., Samoilov V. Direct Mining of Rules from Data with Missing Values. Studies in Computational Intelligence, Chapter in book. Eds. Lin T.Y., Ohsuga S., Liau C.J., Hu X.T., Tsumoto S. Foundation of Data Mining and Knowledge Discovery. Springer. 2005. vol. 6. pp. 233–264.
15. Mill J.S. *Sistema logiki sillogisticheskoy i induktivnoy: Izlozheniye printsipov dokazatel'stva v svyazi s metodami nauchnogo issledovaniya* [System of syllogistic logic and inductive: Statement of principles of proof in relation to the methods of scientific investigation]. Moscow: LENAND. 2011. 832 p. (In Russ.).
16. Five Canons of John Mill. Vikent.ru - portal I.L. Vikent'yeva [Vikent.ru - portal of I.L.Vikent'yev]. Available at: http://vikent.ru/enc/834/ (accessed: 20.06 2014). (In Russ.).
17. Kobylinski L., Walczak K. Efficient Mining of Jumping Emerging Patterns with Occurrence Counts for Classification. Transactions on Rough Sets XIII. LNCS 6499. 2011. pp. 73–88.

18. Sherhod R., Judson P.N., et al. Emerging Pattern Mining To Aid Toxicological Knowledge Discovery. Journal of Chemical Information Modeling. 2014. no. 54(7). pp. 1864–1879.
19. Han J., Pei J., Yin Y. Mining frequent patterns without candidate generation. Proc. of the ACM SIGMOD Intern. Conf. on Management of Data. 2000. pp. 1–12.
20. Blake C.L., Murphy P.M. UCI Repository of machine learning database. University of California, Department of Information and Computer Science. Irvine, CA. 1998. Available at: http://www.cs.uci.edu/mleam/mlrepository.html (accessed: 20.06.2014).
21. Gorodetsky V., Serebryakov S. [Methods and algorithms for collective recognition] *Avtomatika i Telemehanika – Automation and Remote Control*. 2008. no. 11. pp. 3–40. (In Russ.).
22. Gorodetsky V., Samoylov V., Serebryakov S. Ontology–based Context–dependent Personalization Technology. Proc. of the WI/IAT 2010. Toronto. 2010. pp. 278–283.

Тушканова Ольга Николаевна — аспирант, Федеральное государственное бюджетное учреждение науки Санкт-Петербургский институт информатики и автоматизации Российской академии наук. Область научных интересов: машинное обучение, интеллектуальный анализ данных, извлечение знаний, многоагентные системы, рекомендующие системы, облачные технологии, онтологии. Число научных публикаций — 12. tushkanova.on@gmail.com; 14 линия, д. 39, Санкт-Петербург, 199178; р.т.: +79817343119.

Tushkanova Olga Nikolaevna — Ph.D. student, St.Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences. Research interests: data mining, multi-agent systems, recommender systems, cloud computing, ontologies, knowledge extraction technologies. The number of publications — 12. tushkanova.on@gmail.com; 39, 14-th Line, St. Petersburg, 199178, Russia; office phone: +79817343119.

Городецкий Владимир Иванович — д-р техн. наук, заведующий лабораторией интеллектуальных систем, Федеральное государственное бюджетное учреждение науки Санкт-Петербургский институт информатики и автоматизации Российской академии наук. Область научных интересов: искусственный интеллект, технология многоагентных систем, распределенное обучение, извлечение знаний из баз данных, анализ и объединение данных различных источников, P2P сети принятия решений и P2P методы извлечения знаний из данных, обработка больших данных, планирование и составление расписаний, алгоритмы улучшения изображений, рекомендующие системы. Число научных публикаций — 200. gor@mail.iias.spb.su; 14 линия, д. 39, Санкт-Петербург, 199178; р.т.: +7-812-328-3311.

Gorodetski Vladimir Ivanovich — Ph.D., head of laboratory of intelligent systems, St.Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences. Research interests: intelligent data analysis, information fusion, P2P data mining and machine learning, multi-agent systems technology and software tools, agent-based applications, recommender systems, mobile image enhancement. The number of publications — 200. gor@mail.iias.spb.su; 39, 14-th Line, St. Petersburg, 199178, Russia; office phone: +7-812-328-3311.

РЕФЕРАТ

Городецкий В.И., Тушканова О.Н. **Ассоциативная классификация: аналитический обзор. Часть 2.**

В данной работе продолжается рассмотрение методов и алгоритмов ассоциативной классификации. В работе кратко дается постановка задачи ассоциативной классификации. В основной части работы выполнены анализ и сравнение современных моделей, методов и алгоритмов, разработанных в области ассоциативной классификации, основанной на эмерджентных паттернах, применительно к работе с данными большого объема. В заключении формулируются достоинства и недостатки методов ассоциативной классификации и дается оценка перспектив использования этого подхода для интеллектуального анализа больших данных.

SUMMARY

Gorodetsky V., Tushkanova O. **Associative Classification: Analytical Overview. Part 2.**

The paper continues the review of associative classification intended for processing of big data. It shortly formulates corresponding problem statement of associative classification. The main part of the paper represents an overview and comparative analysis of the modern methods, models and algorithms developed for associative classification based on emerging patterns. In conclusion, the paper outlines the main advantages and drawbacks of associative classification, as well as evaluates its capabilities from big data processing perspective.

РУКОВОДСТВО ДЛЯ АВТОРОВ

Взаимодействие автора с редакцией осуществляется через личный кабинет на сайте журнала «Труды СПИИРАН» http://www.proceedings.spiiras.nw.ru. При регистрации авторам рекомендуется заполнить все предложенные поля данных, так как это значительно ускорит процесс оформления метаданных к новым статьям.

Подготовка статьи ведется с помощью текстовых редакторов MS Word 2007 и выше. При подаче материала в редакцию сначала отправляется только статья в формате *.docx. Для обеспечения требований слепого рецензирования при представлении статьи в журнал авторам необходимо удалить персональные данные, содержащиеся в тексте файла и его свойствах.

Объем основного текста – от 5 до 20 страниц включительно. Формат страницы документа – A5 (148 мм ширина, 210 мм высота); ориентация – портретная; все поля – 20 мм. Верхний и нижний колонтитулы страницы – пустые. Основной шрифт документа – Times New Roman, основной кегль (размер) шрифта – 10 pt. Переносы разрешены. Абзацный отступ устанавливается размером в 10 мм. Межстрочный интервал – одинарный. Номера страниц не проставляются.

Не допускается использования цветных шрифтов, цветовых выделений и цветных рисунков. Статьи должны быть полностью готовы к черно-белой печати.

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 материала выделение подразделов.

В основную часть допускается помещать рисунки, таблицы, листинги и формулы. Правила их оформления подробно рассмотрены на нашем сайте в разделе «Руководство для авторов».

ISSN 2078-9181

Вып. 2(39)

ТРУДЫ СПИИРАН
proceedings.spiiras.nw.ru

С-Пб
2015