ISSN 2220-3397

ТЕОРЕТИЧЕСКИЙ И ПРИКЛАДНОЙ НАУЧНО-ТЕХНИЧЕСКИЙ ЖУРНАЛ

Программная инженерия

ПрИн
2
2018
Том 9

Рисунок к статье А. Бернадотт
«АНАЛИЗ НАУЧНОГО ТЕКСТА И НОВЫЕ МИРОВЫЕ ТЕНДЕНЦИИ»

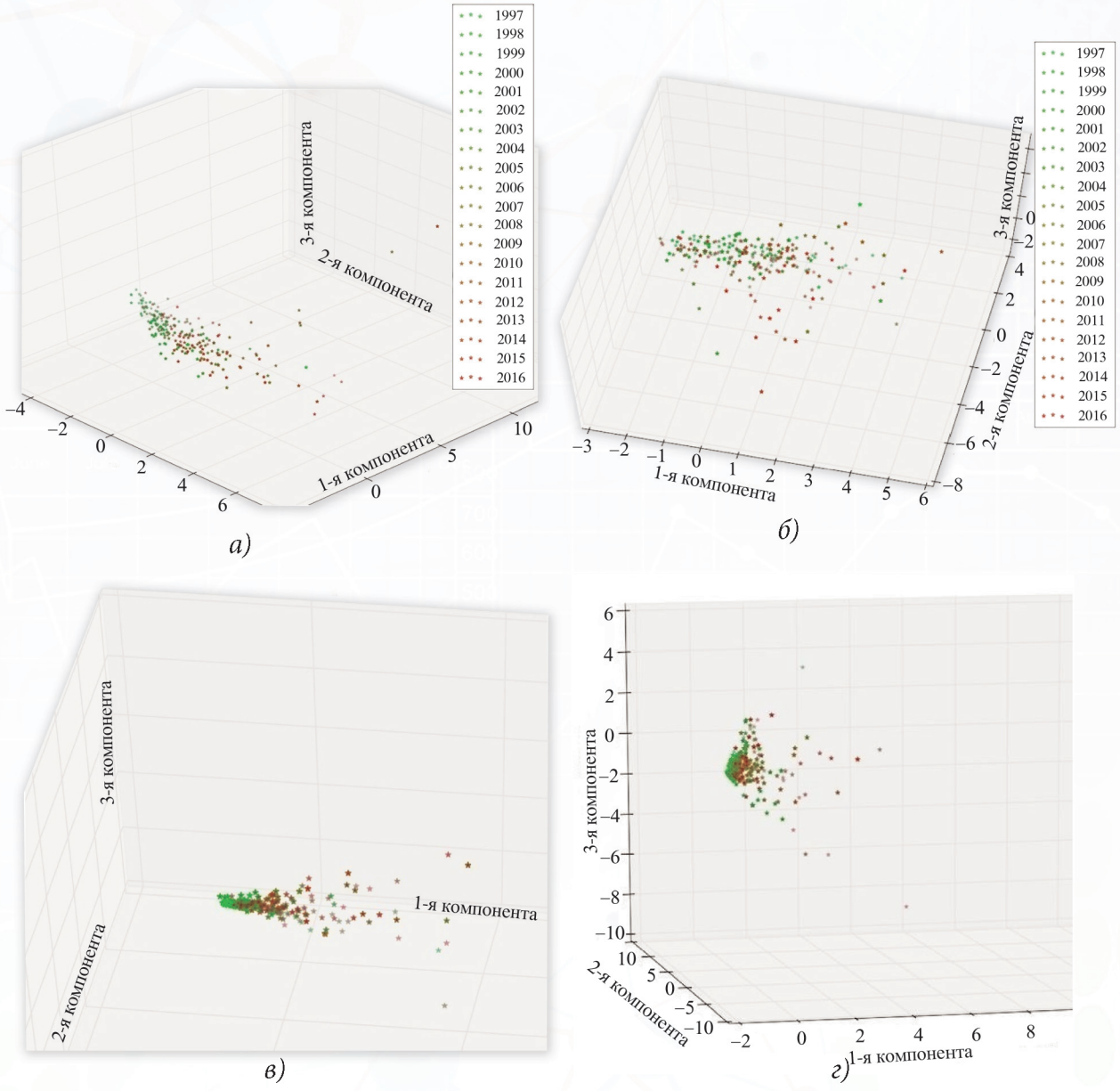

Рис. 2. Анализ главных компонент *tf-idf*-векторов, построенных на:
а – всем корпусе слов; *б* – «фундаментальном» («теоретическом») и «практическом» («прикладном») слово-классах; *в* – «коммерческом» слово-классе;
г – «политическом» слово-классе. Каждой точке соответствует 50 статей

ТЕОРЕТИЧЕСКИЙ И ПРИКЛАДНОЙ НАУЧНО-ТЕХНИЧЕСКИЙ ЖУРНАЛ

Издается с сентября 2010 г. DOI 10.17587/issn.2220-3397 ISSN 2220-3397

Журнал издается при поддержке Отделения математических наук РАН, Отделения нанотехнологий и информационных технологий РАН, МГУ имени М.В. Ломоносова, МГТУ имени Н.Э. Баумана

СОДЕРЖАНИЕ

Журнал зарегистрирован в Федеральной службе по надзору в сфере связи, информационных технологий и массовых коммуникаций. Свидетельство о регистрации ПИ № ФС77-38590 от 24 декабря 2009 г.

Журнал распространяется по подписке, которую можно оформить в любом почтовом отделении (индексы: по каталогу агентства "Роспечать" — **22765**, по Объединенному каталогу "Пресса России" — **39795**) или непосредственно в редакции.
Тел.: (499) 269-53-97. Факс: (499) 269-55-10.
Http://novtex.ru/prin/rus E-mail: prin@novtex.ru
Журнал включен в систему Российского индекса научного цитирования.
Журнал входит в Перечень научных журналов, в которых по рекомендации ВАК РФ должны быть опубликованы научные результаты диссертаций на соискание ученой степени доктора и кандидата наук.

THEORETICAL AND APPLIED SCIENTIFIC AND TECHNICAL JOURNAL

SOFTWARE ENGINEERING

PROGRAMMNAYA INGENERIA

Vol. 9
N 2
2018

Published since September 2010 **DOI 10.17587/issn.2220-3397** **ISSN 2220-3397**

CONTENTS

Information about the jornal is available online at:
http://novtex.ru/prin/eng e-mail:prin@novtex.ru

УДК 001, 004.4 DOI: 10.17587/prin.9.51-58

В. А. Садовничий, академик РАН, проф., ректор, e-mail: info@rector.msu.ru,
В. А. Васенин, д-р физ.-мат. наук, проф., e-mail: vasenin@msu.ru,
МГУ имени М. В. Ломоносова

Интеллектуальная система тематического исследования наукометрических данных: предпосылки создания и методология разработки. Часть 1

В статье, состоящей из двух частей, изложены результаты поисковых и прикладных исследований, направленных на выявление эффективных методов построения и внедрения в практику, сопровождения и развития больших наукометрических систем подготовки и принятия решений в области организации научно-инновационной и образовательной деятельности. Результаты этих исследований получены в ходе проведения предпроектных исследований, на этапах разработки, эксплуатации и развития информационно-аналитической системы (ИАС) "ИСТИНА" (Интеллектуальная Система Тематического Исследования НАукометрических данных) в МГУ имени М. В. Ломоносова.

В первой части статьи изложены предпосылки и краткие сведения по истории создания и становления систем такого назначения, представлены методологические основы их разработки, сопровождения и модификации. Отмечены особенности ИАС "ИСТИНА" как представительной наукометрической системы, учитывающие реалии и специфику России.

Во второй части статьи на примере ИАС "ИСТИНА" будут представлены наиболее значимые, базовые архитектурно-технологические особенности ее реализации, которые на схемах и интерфейсах пользователя проиллюстрируют основные технологические процессы, автоматизируемые или подлежащие автоматизации с помощью этой системы.

Ключевые слова: *наукометрические данные, системы подготовки принятия решений, информационно-аналитическая система, библиометрия, методологические принципы, архитектура, сбор и верификация данных*

Введение

Как и любая сложная, динамичная во времени и открытая к влиянию внешних и внутренних факторов система, научно-техническая и образовательная деятельность без наличия каких-либо регулирующих (управляющих) механизмов предрасположена к хаотичному развитию, непредсказуемо-нежелательной эволюции. Сложность организации и управления процессами на этом направлении обусловлена многообразием участвующих в этой деятельности объектов, субъектов и отношений между ними, которые к тому же динамично меняются. Это обстоятельство является важным, оно отличает науку и образование от других, традиционных направлений в экономической, социальной и политической сферах общественных отношений. Как следствие, такое отличие предъявляет дополнительные требования, накладывает соответствующие им ограничения на подходы, методы и средства управления, которые используются в научно-образовательной среде, усложняет их практическую реализацию.

К числу влияющих на развитие науки и образования внешних факторов, в первую очередь, относятся социальные, экономические и политические. К внутренним факторам можно отнести наличие органического единства и преемственности между различными стадиями образовательного процесса (школа—вуз—подготовка кадров высшей квалификации); состояние издательского дела; возможности обсуждения результатов научно-технической и образовательной деятельности на различных форумах национального и международного уровней; наличие регуляторов кадрового обновления (ротации кадров), а также ряд других.

В связи с тем, что наука и образование являются системообразующими элементами научно-технического прогресса, в значительной степени определяют темпы роста национальной экономики, ведущие государства мира уделяют особое внимание механизмам регулирования этой сферы. Особую значимость на фоне механизмов ее регулирования приобретает и саморегуляция (самоорганизация). К числу базовых механизмов регулирования, как правило, относятся

механизмы стимулирования деятельности, основанные на оценке эффективности ее выполнения. Примерами подобного рода механизмов стимулирования являются: конкурсное избрание и переизбрание на научные и преподавательские должности; оценка результатов работы и оплата за труд; социальные льготы. К составляющим (индикаторам), на которых основываются механизмы стимулирования, относятся:

- количественные и качественные показатели "выпускаемой продукции", включая публикации и патенты, подготовленные кадры различной квалификации и т. п.;
- объемы педагогической деятельности в виде прочитанных курсов лекций и проведенных научных семинаров;
- объемы и качество выполненных научно-исследовательских и опытно-конструкторских работ.

Перечисленные выше механизмы активно использовались в науке и образовании еще в XX веке, однако они имели ряд недостатков. К их числу можно отнести: ограниченный перечень индикаторов деятельности; статический характер процедур определения их значений (как правило, один раз в конце года); отсутствие возможности оперативного и объективного сравнительного анализа этих данных применительно к отдельному ученому или педагогу.

Эти недостатки, как показывают результаты научных исследований (1960—1970 гг.) в области наукометрии, приводят к появлению в сфере науки и образования негативных процессов (публикации, научные форумы, инновационные показатели и т. п.), которые именуются [1] "адаптированным торможением" (самоторможением). Такие процессы приводят к созданию препятствий (барьеров) для распространения новых идей, активной инновационной деятельности, а также к другим деструктивным явлениям, которые на практике должны способствовать научно-техническому прогрессу. Как показывают реальная действительность и результаты исследований, такие явления происходят даже несмотря на благоприятные макроэкономические факторы и применяемые стимулы, такие как рост оплаты труда, увеличение числа научно-педагогических работников, социальных льгот и т. п. Аналогичные явления наблюдаются и в сфере образования. Это вполне естественно, так как традиционно в мире значительная часть научных исследований, а также подготовка кадров для них осуществляются в университетах, которые через отдельных преподавателей, ведущих исследовательскую деятельность, связаны с научными центрами.

Первая и, пожалуй, главная причина самоторможения в науке и, как следствие, в образовании обусловлена тем обстоятельством, что значительная часть ученых, исследовательских групп (научных школ и др.), которые выросли на некогда новых идеях и добились определенных успехов с применением методов и средств реализации этих идей и используя сложившийся научный авторитет, начинают препятствовать вновь появляющимся идеям и подходам, их объективному обсуждению, адекватной оценке и распространению. Второй важной причиной может быть с годами складывающаяся возрастная диспропорция (дисбаланс) в исследовательском и/или преподавательском коллективе, отсутствие в нем действенных механизмов, способствующих научному росту (продвижению) молодых его членов. Существуют и другие, менее значимые причины подобных явлений.

Отмеченные выше причины и порождающие их обстоятельства приводят к деструктивным воздействиям на научно-образовательную среду (сферу деятельности, инфраструктуру сопровождения и т. п.) и, как следствие, приводят к непредсказуемому (нежелательному, хаотичному) ее поведению (изменению, эволюции). Здесь следует отметить, что такие изменения, к счастью, происходят медленно во времени, что обусловлено консервативно-инерционным характером этой сферы деятельности. В качестве примера подобных нежелательных изменений можно привести известные деградационные процессы в сфере науки и образования России, которые наблюдаются с 1980-х гг. по настоящее время и, как следствие, сложившуюся на сегодня ситуацию в РАН, в российском высшем образовании.

Для устранения отмеченных выше недостатков, присущих традиционно используемым механизмам оценки эффективности и стимулирования деятельности в сфере науки и образования, необходимы более тонкие, ориентированные на процессы саморегулирования механизмы. Понимание важности решения такой задачи пришло к ученым-исследователям в области наукометрии к началу 1980-х гг. К этому времени в мире сложились уже и технологические предпосылки к ее решению в виде активно развивающихся средств вычислительной техники, а затем и средств связи на основе технологии пакетной коммутации (источников метасети Интернет). На этой технологической базе появилась возможность не только увеличить число субъектов, объектов, отдельных показателей научно-технической деятельности, но и сопровождающих эту деятельность отношений между перечисленными сущностями. Представилась возможность оперативной их обработки.

На новой технологической основе стали формироваться новые подходы к традиционной библиометрии, способы оценки эффективности научной и инновационной деятельности, состояния разработки и внедрения в практику наукоемкой продукции. В качестве информационной базы для их реализации крупными издательскими корпорациями в начале 2000-х гг. стали создаваться центры индексирования и анализа библиометрических данных. К числу самых больших таких центров применительно к англоязычным публикациям относятся Web of Science компании Thomson Reuters и Scopus компании Elsevier, применительно к русскоязычным публикациям — РИНЦ (Российский Индекс Научного Цитирования).

На базе информационных коллекций таких центров стали создаваться инструментальные средства анализа состояния и оценки эффективности деятельности отдельных субъектов научно-технической деятельности. К числу таких средств относятся PURE, SciVal/Elsevier, Converis/Thomson Reuters. Такого сорта средства принято именовать *Current Research Information System* — CRIS.

Отмеченные выше CRIS-системы включают программные механизмы (далее — механизмы), позволяющие по заранее заданным в запросе критериям (показателям) оценить не только результаты деятельности организаций и отдельных научных коллективов, но и ученых индивидуально. Однако они имеют ряд недостатков, которые ограничивают возможности их использования. В современных, активно используемых на практике CRIS-системах отсутствуют или не в должной мере представлены следующие механизмы, которые позволяют реализовать перечисленные далее процессы, востребованные для подготовки принятия управленческих решений на разных уровнях управления наукой и образованием.

- Механизмы, поддерживающие оперативный характер поступления, верификации и анализа разноплановых сведений о результатах персональной деятельности (ученых, педагогов, инженерно-технических работников), которые могут использоваться при премировании, в различного рода конкурсных процедурах и в других стимулирующих действиях.
- Оценка составляющей результатов деятельности по подготовке научных кадров на направлении исследования объекта, имеющего хорошие инновационные перспективы.
- Оценка степени эффективности участия отдельных работников в коллективно выполняемых НИР и НИОКР.
- Учет эффективности научно-образовательной деятельности коллективов на разных уровнях (научная группа, кафедра/лаборатория, вуз/НИИ).
- Логическое разграничение доступа объектов к различным сервисам и данным в условиях перманентного (интенсивного во времени) изменения субъектов, объектов и отношений между ними.
- Учет и анализ эффективности использования уникального оборудования в научных исследованиях и в образовательном процессе.

CRIS-системы, имеющие международное признание, в большей степени ориентированы на цели исследований и реалии, которые характерны для зарубежных стран, и основаны, как правило, на учете и индексировании англоязычных публикаций.

Недостатки существующих CRIS-систем привели к необходимости разработки в МГУ имени М. В. Ломоносова информационно-аналитической системы (ИАС) "ИСТИНА" (Интеллектуальная Система Тематического Исследования НАукометрических данных) [2]. Ключевыми целями этой разработки были желание устранить отмеченные выше недостатки и ее ориентация на реалии, интересы и потребности России и других стран, активно использующих русский язык.

Интеллектуальный характер ИАС "ИСТИНА" (далее — Система) обусловлен:

- наличием в ее составе формальных моделей и реализующих их программных механизмов, которые позволяют аккумулировать уже существующие (сложившиеся) знания (фактографические сведения) о проблемной области (наукометрии) и пополнять их по мере появления новых;
- возможностью использовать эти знания для интеллектуального анализа состояния проблемной области на разных уровнях ее описания (от персонального и коллективного до регионального и национального) для получения данных по запросам пользователей.

Регулирование процессов, сопровождающих научную и образовательную деятельность, в Системе осуществляется путем подготовки исходных данных (сведений) с помощью имеющихся в ее распоряжении инструментальных средств и административных механизмов. На основе таких данных и последующей экспертизы коллегией экспертов принимаются социальные, а также экономически и политически мотивированные решения. Эти решения стимулируют на конкурентной основе к более эффективной деятельности ученых и педагогов, научные коллективы и организации. Элементы саморегулирования в рамках такого подхода к управлению поддерживаются программными механизмами Системы. Такие механизмы позволяют каждому субъекту деятельности (отдельному работнику или коллективу лаборатории/кафедры) сравнить оценку эффективности своей деятельности с усредненной оценкой по структурному подразделению вышележащего уровня (факультет, институт, организация в целом). На этой основе каждый такой субъект видит сформированную в предварительном порядке и коллегиально принятую оценку своей профессиональной деятельности и в качестве ответной реакции может спланировать меры и оперативные действия по ее поддержке на надлежащем уровне.

1. Основные принципы и требования к ИАС "ИСТИНА"

Отметим одно важное обстоятельство, которым изначально руководствовались разработчики ИАС "ИСТИНА". Под наукометрией здесь и далее понимаются не только модели и методы оценки научно-инновационной деятельности, но и работы по подготовке кадров, осуществляющих эту деятельность. Такой подход к подготовке кадров специалистов на основе их активного участия в научных исследованиях, который принято называть "гумбольдтовским", используется во многих ведущих университетах мира, в том числе — в Московском государственном университете.

1.1. Базовые принципы создания и развития

К базовым принципам создания и развития ИАС "ИСТИНА" относятся перечисленные далее основополагающие постулаты.

- Целевые установки Системы должны быть: направлены на стимулирование научно-инновационной и образовательной деятельности отдельных работников, коллективов и организаций; основаны на разумном анализе, учете и балансе текущих интересов и особенностей национального, регионального и корпоративно-локального характера.
- Методология создания и развития Системы должна опираться на единые на всех уровнях организации науки и образования: представления о ее целевых установках; описание модели наукометрии как проблемной области, имея в виду составляющие ее сущности (субъекты, объекты и отношения между ними).
- Процессы формирования фактографической базы исследования в приоритетном порядке должны отражать восходящий ("снизу вверх") принцип его организации, который сочетает интересы отдельных работников и стимулы к их эффективной работе на всех уровнях организации научной и образовательной деятельности (постулат универсален, в том числе — соответствует предыдущему).

1.2. Методологические принципы и требования к Системе

В основе методологических принципов и соответствующих им требований к Системе лежат положения Лейденского манифеста для наукометрии, принятые международным сообществом в 2014 г. [3]. Эти положения достаточно подробно обсуждались в публикациях, на различных международных форумах и относительно просты для восприятия. В связи с этим не будем их комментировать подробно.

1. Количественная оценка деятельности отдельного работника является лишь отправной для последующей ее оценки коллегией экспертов.

2. Подлежащая оценке научная и образовательная деятельность и соответствующие индикаторы должны выбираться с учетом целевых установок и особенностей ее организации в тех или иных областях знания.

3. Методы оценки деятельности должны учитывать национальные и региональные особенности, в первую очередь — целевые установки ее организации.

4. Процессы сбора данных и их анализа должны быть открытыми, прозрачными и простыми для понимания и возможности их воспроизведения.

5. В Системе должны присутствовать механизмы, предоставляющие возможности работникам, деятельность которых оценивается, проверять и анализировать данные, которые положены в основу оценки.

6. Методы оценки должны учитывать тот факт, что научные дисциплины отличаются друг от друга по практике публикаций и цитирования.

7. Количественная оценка отдельных работников должна учитывать их индивидуальные особенности, включая возраст, стаж работы, особенности области знания и др.

8. Методы оценки должны быть избавлены от индикаторов, незначительно влияющих на целевые установки и конечный результат.

9. Стимулирующее воздействие оценки должно соответствовать целевым установкам, сложившимся на настоящее время.

10. Проверка индикаторов на их соответствие целям и задачам Системы должна быть перманентной.

Не вдаваясь в подробности отметим, что в настоящее время все перечисленные принципы в Системе успешно реализуются. Более подробная информация, подтверждающая это, будет представлена во второй части статьи.

Учитывая реалии развития наукометрии в России, разработчики ИАС "ИСТИНА" руководствуются также и рядом других методологических принципов, которые отражают такие реалии. Отметим их и кратко прокомментируем далее.

11. При сборе и верификации данных в больших наукометрических коллекциях предпочтение следует отдавать восходящим потокам их поступления, основанным на персональном или коллективном интересе источника данных.

В отсутствие достаточно больших объемов хорошо верифицируемых данных, определяющих значения индикаторов (которые могут содержаться в наукометрических базах, например, WoS, Scopus и др.), в условиях изменения самих индикаторов и "формул оценки", собрать "критический" объем таких данных — это отдельная и очень сложная задача. Под "критическим" здесь будем иметь в виду объем данных в Системе, который позволяет на его основе запускать механизмы стимулирования научно-технической и преподавательской деятельности на всех уровнях деятельности отдельной организации.

В отличие от зарубежных CRIS-систем, которые изначально базируются на большом объеме уже собранных данных в библиометрических базах, создание Системы в России лишено такой возможности. В связи с этим наиболее эффективным представляется восходящий ("снизу вверх") от отдельной персоны и верифицируемый сбор таких данных.

12. Процессы сбора и верификации данных должны рационально сочетать не только приоритетные восходящие "снизу вверх", но и нисходящие "сверху вниз" (из агрегированных данных к составляющим их персональным) потоки.

Первые при этом, как правило, лучше поддаются верификации и быстрее создают актуальную "критическую" массу коллекции, результаты исследований на которой обладают большей степенью доверия.

Отдавая, согласно этому принципу, приоритет процессам "снизу вверх", нельзя отказываться от потоков сбора данных "сверху вниз". При этом данные, полученные с помощью сбора "сверху вниз", должны быть подвергнуты дополнительной верификации с использованием доступных (принятых) в Системе механизмов. В отсутствие должной верификации коллекция достаточно быстро может оказаться "загрязненной" и, как следствие, неактуальной. Это

вызовет нарекания (критику) и неприятие Системы пользователями и остановит ее развитие.

13. Процессы сбора и верификации должны основываться на рациональном (сбалансированном) сочетании для отдельных работников и для коллективов стимулов как принудительного (приказного, обязывающего) характера, так и стимулов, основанных на персональном и/или коллективном интересе.

Стимулы, основанные на персональном и коллективном интересе (конкурсы, премии, социальные стимулы) и т. п., — действенное средство повышения эффективности деятельности организации. Они, как правило, являются и самыми действенными рычагами, ускоряющими процессы развития Системы, повышающими ее авторитет в среде потенциальных пользователей.

14. Методы и, соответственно, индикаторы и метрики оценки эффективности результатов деятельности научного работника должны в сбалансированном режиме включать и составляющие, характеризующие результаты его деятельности в подготовке кадров специалистов в соответствующей области науки, и наоборот, для педагогов — результаты их научной деятельности.

В противном случае, даже при наличии успехов в научной деятельности как отдельных ученых, так и коллективов, будет теряться подготовка научных кадров, способных "продвигать" эту область науки. Более того, "перекосы" в оценке эффективности могут приводить к стремлению научного работника заниматься учебной деятельностью по совместительству в другой организации, и наоборот.

15. Методы оценки эффективности результатов субъектов научной деятельности (и персональной, и коллективов) должны должным образом учитывать ее инновационную составляющую.

В противном случае стремление к публикациям в престижных, в первую очередь, зарубежных журналах будет работать в ущерб получению востребованных на практике, в том числе национально значимых, стратегических решений.

16. При соблюдении принципов прозрачности (транспарентности) данных общего характера для "широкого" научно-образовательного сообщества в соответствии с положениями нормативно-законодательной базы РФ механизмы Системы должны гарантировать:

- **каждому ученому и педагогу конфиденциальный статус (конфиденциальность) данных персонального характера, а также информации, которая составляет оценку его личной профессиональной деятельности;**
- **каждой организации и органу государственного управления конфиденциальность агрегированных данных, которые не подлежат разглашению, в том числе — положениями документов локального и ведомственного уровней.**

В ИАС "ИСТИНА" программно реализованы современные, имеющие все признаки интеллектуальной новизны модели логического разграничения доступа к информационным ресурсам, учитывающие возможности ее использования на всех уровнях административной иерархии (структурного подразделения, организации в целом, региональный и национальный уровни) в режиме распределенной BigData-системы, функционирующей как традиционные социальные сети.

17. Механизмы и данные в Системе в первоочередном порядке должны быть ориентированы на учет национальных особенностей, интересов и стратегических целевых установок на развитие научного и образовательного потенциала России, а также вытекающих отсюда задач каждой отдельной организации.

Показатели, оценивающие эффективность деятельности структурных подразделений МГУ в ИАС "ИСТИНА", ориентированы на официально объявленные стратегические направления развития науки и образования в России. К их числу относятся:

- приоритетные направления развития науки, технологий и техники РФ;
- критические технологии РФ;
- приоритетные направления развития МГУ до 2020 г.;
- стратегические направления в российском образовании.

Как следствие, поскольку организация информации в ИАС "ИСТИНА" идет преимущественно "снизу вверх" — от персон до структурных подразделений, то индикаторы, оценивающие эффективность работы отдельных работников, стимулируют деятельность коллективов (кафедр, лабораторий, вузов и НИИ), направленную на реализацию положений этих документов.

18. В основе методологии создания, становления, сопровождения (эксплуатации) и развития Системы должен лежать принцип, позволяющий одновременно:

- создавать математическое, алгоритмическое и программное обеспечение отдельных функционально замкнутых, новых компонентов и блоков Системы;
- обеспечивать доведение вновь вводимых компонентов и блоков Системы до состояния их функциональной самодостаточности и востребованности на практике;
- сопровождать Систему в процессе решения практических задач, на основе анализа результатов эксплуатации обеспечивать ее модернизацию (рефакторинг, реинжиниринг) без потери уже присутствующих функциональных возможностей.

Методология создания и развития ИАС "ИСТИНА" изначально ориентирована и обеспечивает деятельность в рамках проекта, которая предусматривает одновременное выполнение работ на всех перечисленных направлениях. Необходимость реализации такого подхода к созданию и развитию Системы обусловлена тем обстоятельством, что не существуют и объективно не могут быть определены требования к Системе, необходимые для их реализации применительно к целевой задаче. Причины в новизне, оригинальности и наукоемком характере целевых установок на создание Системы с такими свойствами,

как следствие — отсутствие даже канонизированных требований к ней.

Все перечисленные выше методологические принципы, положенные в основу ИАС "ИСТИНА", реализуются в рамках Проекта "Развитие и сопровождение информационно аналитической системы подготовки принятия решений на основе анализа информации о результатах научно-исследовательской, педагогической и инновационной деятельности ИАС "ИСТИНА" (этап 1 — 2013 г., этап 2 — 2013—2014 гг., этап 3 — 2016—2017 гг.).

1.3. Архитектурные принципы и требования к Системе

Архитектура Системы, отвечающая представленным ранее базовым и методологическим принципам ее создания и перманентного развития, модернизации и сопровождения по назначению, должна, по мнению разработчиков ИАС "ИСТИНА", основываться на перечисленных далее принципах.

1. Архитектура Системы должна быть модульной и масштабируемой как на макроуровне ее описания, так и на нижележащих уровнях, и адекватно отразить процессы, подлежащие автоматизации.

2. Архитектура Системы должна быть иерархически организована (структурирована) и процессно-ориентирована на каждом из уровней структурной иерархии, включая следующие блоки на отдельных уровнях:

- функционально-замкнутые блоки (компонентов, модулей) Системы, ориентированные на автоматизированную реализацию макропроцессов, реализующих функции в рамках этого блока;
- функционально-обособленные компоненты (модули) в составе отдельных блоков Системы, ориентированные на реализацию соответствующих этому модулю функций (процессов);
- компонент Системы, реализующий функции монитора безопасности для разграничения доступа к отдельным модулям Системы, к базам данных и отдельным данным в таких базах;
- компонент Системы, реализующий связные (передачи данных) и интегрирующие функции в Системе.

3. Архитектура Системы должна с достаточной полнотой отражать все автоматизируемые в рамках Системы процессы, которые востребованы практикой наукометрии, а также задачами в настоящем и на прогнозируемую перспективу в области подготовки к принятию в этой области управленческих решений.

4. Архитектура Системы должна учитывать процессы взаимодействия (обмена данными, запросы и т. п.) с базами различного рода вспомогательных данных, дополняющих, актуализирующих и конкретизирующих (уточняющих) данные, которыми располагает собственно Система, имея в виду как внутренние, так и внешние по отношению к организации базы.

При разработке, развитии и модернизации ИАС "ИСТИНА", при ее сопровождении по назначению соблюдаются перечисленные выше архитектурные принципы и вытекающие из них общие требования к Системе. Они реализуются в том объеме, которому отвечает текущий уровень реализации настоящего проекта.

1.4. Технологические принципы и требования

Далее на примере ИАС "ИСТИНА" кратко сформулируем технологические принципы создания и развития Систем, аналогичных ей по назначению и условиям использования.

1. Соблюдение положений нормативных документов РФ применительно к созданию, эксплуатации и развитию национально значимых систем, в том числе имея в виду перспективы использования отечественного программного обеспечения.

Эти требования пока не реализованы в полной мере. Причина в том, что пока нет готовых к полномасштабному использованию отечественных программно-аппаратных средств или систем с открытым программным кодом, которые можно использовать для разработки подобных Систем, больших по объемам задействованных в них сущностей со сложными отношениями между ними, по их системной поддержке, по механизмам управления и данными, и знаниями о предметной области (наукометрия). Это существенно ограничивает возможности реализации отмеченного принципа в полном объеме. Однако в ближайших планах — деятельность по реинжинирингу Системы, устраняющему недостатки на этом направлении.

2. Соблюдение основных положений и рекомендаций к инженерии программ на всех этапах жизненного цикла Системы.

В настоящее время в жизненном цикле отдельных компонентов ИАС "ИСТИНА" предусмотрены и реализуются: предпроектные исследования; оценка эффективности предлагаемых решений; проектирование; программная реализация; тестирование и т. п. Каждому ресурсоемкому компоненту в ИАС "ИСТИНА" и в Системе в целом соответствуют: модели; алгоритмы; ЕСПД-документация; код с комментариями.

3. При вторичной (по запросу) обработке данных Система должна в максимально возможной степени использовать механизмы взаимодействия с базами данных и содержащимися в них данными (в том числе — с точки зрения их конфиденциальности), которые необходимы для выполнения запроса.

Это требование позволит активно использовать данные как де-факто существующих баз данных, так и вновь разрабатываемых без ущерба для конфиденциальности части данных в этих базах. Здесь следует отметить то обстоятельство, что существующие и используемые на практике модели и программные средства обеспечения информационной безопасности (в первую очередь — разграничения доступа к ресурсам) таких сложно организованных Систем не могут удовлетворить требованиям, которые следуют из положений этого принципа. Однако определенные результаты исследований и практической

реализации на этом направлении есть, и они будут представлены во второй части статьи.

4. Механизмы (математическое, алгоритмическое и программное обеспечение) должны учитывать различные уровни конфиденциальности данных, которыми располагает Система и те базы данных, к которым она может обращаться.

Механизмы логического разграничения доступа к данным в ИАС "ИСТИНА" опираются на формальные модели их описания, которые отличаются от традиционно принятых в "классической" информационной безопасности. Они аккумулируют мировой опыт создания такого сорта механизмов в социальных сетях и учитывают особенности наукометрии как проблемной области.

5. Система должна поддерживать интеграционные механизмы, позволяющие извлекать данные, необходимые по запросу пользователей, из других, в том числе удаленных в сети Интернет баз данных (БД), с их защитой от несанкционированного доступа и верификацией, с установлением их соответствия ("привязкой") к отдельным персонам из БД Системы.

В настоящее время с той или иной степенью завершенности в ИАС "ИСТИНА" реализованы следующие механизмы интеграции данных:

- механизмы обмена информацией по телекоммуникационным каналам между пользователями Системы и базами данных (БД, внутренними и внешними по отношению к Системе) с использованием единой модели логического разграничения доступа (МЛРД) к ресурсам Системы;
- механизмы ввода данных в Систему в режимах "снизу вверх" (от конечного пользователя) и "сверху вниз" (из ранее сформированных источников) с их первичной обработкой, верификацией, "привязкой" к персонам и размещением в БД Системы под контролем МЛРД;
- механизмы оперативного вывода результатов вторичной обработки и агрегирования данных по запросам пользователей Системы (включая отдельных работников, а также ответственных за сопровождение информации в ИАС "ИСТИНА" от структурных подразделений и органов управления организации в целом) под контролем программной реализации МЛРД.

1.5. Математическая модель наукометрии

Принимая во внимание отмеченные выше принципы создания и развития наукометрических систем на примере ИАС "ИСТИНА", возникает естественный вопрос о возможностях построения формальной модели описания наукометрии как отдельной проблемной области. Далее кратко представим соображения по целесообразности и по подходам к разработке такой модели, а также по возможности ее использования в процессах создания, развития и сопровождения Системы, в процессах модификации.

Математическая модель наукометрии может быть создана на основе:

1) простого отображения (рейтинговая формула) точки в заранее принятом n-мерном пространстве R_n индикаторов в оценку на R_1;

2) создания онтологий проблемной области, включая систему характеризующих проблемную область понятий и классов, экземпляров объектов, отношений между ними, способов интеграции таких характеристик и их отображения в числовые показатели;

3) построения теоретико-множественной модели проблемной области в виде базы знаний;

4) построения моделей на основании уже существующих феноменологических моделей в других проблемных областях;

5) построения многоагентной модели.

В основу формальной модели наукометрии, которая в настоящее время используется в ИАС "ИСТИНА", положена вторая из перечисленных выше. Такая модель наукометрии может обеспечивать решение перечисленных далее задач.

- Верификация программного кода при его модификации.
- Решение уже востребованных практикой задач:
  - нормализация оценочных рейтинговых показателей для различных областей знаний;
  - обеспечение приоритетов "определяющих" индикаторов над "второстепенными".
- Поддержка эффективных механизмов поиска, кластеризации и анализа данных по различным тематическим запросам.
- Получение оценки сложности программных реализаций (вычислений) различных запросов.
- Оценка тенденций научных исследований и образования во взаимодействии с математическими моделями их описания (феноменологические, теоретико-вероятностные, многоагентные и другие подходы).

Заключение

Представленные в настоящей статье базовые методологические, архитектурные и технологические принципы создания, сопровождения и развития ИАС "ИСТИНА", а также вытекающие из них требования к этой системе, являются, по мнению авторов, общими для систем наукометрии, учитывающих социально-экономические, политические (внутренние и геополитические) реалии России, и направленных на решение возникающих в связи с ними стратегических задач в области науки и образования. Эти принципы и требования могут быть использованы при построении других систем аналогичного назначения. На базе таких систем в перспективе должно сформироваться единое в масштабах страны, методологически однородное по подходам к их реализации поле, позволяющее осуществлять перманентный контроль (мониторинг) состояния науки и образования в России.

Во второй части статьи, которая дополнит содержание настоящей, будет показано, как эти основополагающие принципы и требования реализованы на практике в процессе создания и развития ИАС "ИСТИНА".

Список литературы

1. **Налимов В. В., Мульченко З. М.** Наукометрия. Изучение развития науки как информационного процесса. Физико-математическая библиотека инженера. М.: Наука, 1969. 192 с.

2. **Интеллектуальная** система тематического исследования научно-технической информации (ИСТИНА) / Под ред. В. А. Садовничего. М.: Изд-во МГУ, 2014, 262 с.

3. **Хикс Д., Воутерс П., Волтман Л.** и др. (пер. А. А. Исэрова) Лейденский манифест для наукометрии // Нанотехнологическое Сообщество. 2015. URL: http://www.nanometer.ru/2015/07/31/naukometria_464938.html

Intellectual System of Thematic Investigation of Scientometrical Data: Background of Creation and Methodology of Development

V. A. Sadovnichy, info@rector.msu.ru, **V. A. Vasenin,** vasenin@msu.ru, Lomonosov Moscow State University, Moscow, 119991, Russian Federation

Corresponding author:

Vasenin Valery A., Professor, Lomonosov Moscow State University, Moscow, 119991, Russian Federation
E-mail: vasenin@msu.ru

Received on November 16, 2017
Accepted on December 07, 2017

In the paper, the contents of which will be presented in two parts, the results of research and applicable investigations are presented, which are aimed at designing and introducing into practice, maintaining and developing large scientometrical systems for preparing and making decisions in the realm of organizing scientific, innovational and educational activity. The results of these investigations were obtained during developing, exploiting and introducing the "ISTINA" information analysis system (Intellectual System of Thematic Investigation of Scientometrical Data) in Lomonosov Moscow State University, which is aimed at solving the problems of this kind.

In the first part the background and brief information on creation and development of systems of such designation is given, the methodology of developing, maintaining and modifying them is presented. Applied to IAS "ISTINA" as a representative scientometrical system, its features which take the realities and specifics of Russia into account are mentioned.

In the second part based on the example of IAS "ISTINA" the most important, basic architectural and technological specifics of its implementation will be presented, which will illustrate on schemes and user interfaces the main technological processes that are being automated or ought to be automated using this system.

Keywords: *scientometrical data, systems of preparing decisions, information analysis system, bibliometrics, methodological principles, architecture, collecting and verifying data*

For citation:

Sadovnichy V. A., Vasenin V. A. Intellectual System of Thematic Investigation of Scientometrical Data: Background of Creation and Methodology of Development, *Programmnaya Ingeneria,* 2018, vol. 9, no. 2, pp. 51—58

DOI: 10.17587/prin.9.51-58

References

1. **Nalimov V. V., Mul'chenko Z. M.** *Naukometrija. Izuchenie razvitija nauki kak informacionnogo processa. Fiziko-matematicheskaja biblioteka inzhenera* (Scientometrics. A study of the development of science as an information process. Physico-mathematical library engineer), Moscow, Nauka, 1969, 192 p. (in Russian).

2. **Intellektual'naja** *sistema tematicheskogo issledovanija nauchno-tehnicheskoj informacii* (ISTINA) (Intelligent system case studies of scientific and technical information (ISTINA)) / Ed. V. A. Sadovnichy, Moscow, Izd-vo MGU, 2014, 262 p. (in Russian).

3. **Hicks D., Wouters P., Waltman L., de Rijcke S., Rafols I.** Bibliometrics: The Leiden Manifesto for research metrics, *Nature,* 2015, vol. 520, pp. 429—431.

УДК 004.4, 004.94, 004.7 DOI: 10.17587/prin.9.59-68

Д. А. Басавин[1], ассистент кафедры, e-mail: basavind@gmail.com,
С. В. Поршнев[2], д-р техн. наук, проф., e-mail: sergey_porshnev@mail.ru,
Д. А. Петросов[1], канд. техн. наук, доц., зав. кафедрой, e-mail: scorpionss2002@mail.ru,
[1] Белгородский государственный аграрный университет им. В. Я. Горина, п. Майский, Белгородская обл.,
[2] Уральский федеральный университет имени первого Президента России Б. Н. Ельцина, г. Екатеринбург

Сравнение последовательной и параллельной программных реализаций гибридной жидкостной модели информационных потоков для компьютерных сетей со сложной топологией

Описаны результаты сравнения последовательной, использующей только центральный процессор (Central Processing Unit, CPU) и параллельной, использующей графический процессор (Graphic Processing Unit, GPU), программных реализаций гибридной жидкостной модели информационных потоков в компьютерных сетях со сложными топологиями. Показано, что рассчитанные с помощью последовательной и параллельной программных реализаций гибридной жидкостной модели количественные характеристики информационных потоков согласуются друг с другом. Проведено графическое и численное сравнение следующих зависящих от времени агрегированных на входах и выходах маршрутизаторов количественных характеристик: средняя нагрузка, максимальная нагрузка, минимальная нагрузка. В ходе исследования были получены оценки зависимости времени вычислений от числа информационных потоков и маршрутизаторов, свидетельствующие о целесообразности использования параллельной реализации гибридной жидкостной модели при числе моделируемых информационных потоков более 2^6.

Ключевые слова: *интернет-трафик, компьютерные сети, параллельная гибридная жидкостная модель, имитационное моделирование, графические процессоры общего назначения, технология GPGPU*

Введение

В настоящее время активно осуществляется переход на так называемые безлимитные тарифные планы. В них одним из основных критериев качества обслуживания является предоставление заявленной пользователю скорости передачи данных. В связи с этим резко возрастают требования к техническим решениям, принимаемым при проектировании современных телекоммуникационных сетей. При этом одним из основных вопросов, который волнует операторов, является вопрос о числе пользователей, которое можно обслуживать при условии обеспечения заявленной скорости доступа к интернет-каналам на имеющемся в наличии сетевом оборудовании. Однако, как показывает практика, в настоящее время по-прежнему наиболее популярными остаются эмпирические методы оценки технических характеристик сетевого оборудования (например, метод экспертных оценок), которые в ряде случаев оказываются недостаточно точными. Такой подход, в свою очередь, может приводить как к неоправданным затратам, так и к несоответствию между заявленным и фактическим качеством предоставляемых пользователю сетевых сервисов и, как следствие, невыполнению провайдером взятых на себя обязательств.

Анализ известных решений отмеченной задачи показывает, что подход, основанный на использовании математических моделей информационных потоков в компьютерных сетях (КС), позволяет сделать адекватный выбор необходимого сетевого оборудования и при этом не требует его приобретения, монтажа, настройки и проверочного тестирования на действующих интернет-каналах. Среди известных математических моделей информационных потоков в КС можно выделить перечисленные далее классы, которым, однако, присущи известные недостатки [1].

1. Аналитические модели, к которым, в первую очередь, относятся модели теории массового обслуживания. В данном классе моделей не удается учесть особенности современных механизмов регулирования скорости. К их числу относятся механизмы реакции на потери пакетов, ограничения потоков отдельных пользователей, влияние потоков друг на друга и т. д. [2, 3].

2. Программы — генераторы сетевого трафика, к которым относятся статистические модели трафика. В данном классе моделей отсутствуют возможности учета особенностей передачи генерируемого трафика по каналам передачи данных и возможности реализации механизмов обратной связи при потере пакетов. Механизмы обратной связи практически повсеместно используются в современных интернет-каналах для управления скоростью передаваемого потока данных.

3. Сетевые пакетные симуляторы — специализированные программные продукты, предназначенные для моделирования каналов со средней (порядка десятков Мбит/с) пропускной способностью. Данные модели позволяют описать процесс передачи данных по сети на уровне отдельных пакетов и учесть механизмы регулирования скорости потоков трафика.

4. Жидкостные модели (ЖМ), учитывающие механизмы управления скоростью потоков передачи. Такие механизмы позволяют существенно уменьшить число рассматриваемых событий при моделировании интернет-трафика за счет перехода от рассмотрения процессов распространения в канале передачи данных отдельных пакетов к рассмотрению укрупненных групп пакетов.

Напомним, что в ЖМ сетевой трафик описывается в терминах изменения во времени скорости передачи данных i-го потока $W_i(t)$ и длины очереди на входе в l-й канал $q_l(t)$. Данная модель также учитывает потерю пакетов при их передаче и позволяет реализовывать современные алгоритмы управления скоростью информационных потоков. Жидкостная модель представляет собой следующую систему дифференциальных уравнений (СДУ):

$$\frac{dW_i(t)}{dt} = \frac{1\left(W_i(t) < M_i\right)}{R_i(t)} - \frac{W_i(t)}{2}\lambda_i(t), \qquad (1)$$

$$\frac{dq_l(t)}{dt} = -1\left(q_l(t) > 0\right)C_l + \sum_{i \in N_l} n_i A_i^l(t). \qquad (2)$$

В этих уравнениях $1_{f(t)}$ — функция Хевисайда:

$$1_{f(t)} = \begin{cases} 1, & f(t) \geqslant 0, \\ 0, & f(t) < 0, \end{cases}$$

$R_i(t)$ — время двойного оборота (RTT) i-го потока; $\lambda_i(t)$ — скорость потери пакетов i-го потока; C_l — пропускная способность l-го канала, который обслуживается данным маршрутизатором; $A_i^l(t)$ — скорость пребывания пакетов по l-му каналу, которая рассчитывается следующим образом:

$$A_i^l = \begin{cases} A_i(t), & l = k_{i,1}, \\ D_i^{b(l)}\left(t - a_{b_i(l)}\right), & l \neq k_{i,1}, \end{cases}$$

где $k_{i,1}$ — первая по счету очередь в маршруте следования пакетов соответствующего информационного потока; $b_i(l)$ — предшествующая очередь в маршруте следования пакетов соответствующего информационного потока; a — задержка распространения сигнала по соответствующему каналу связи; $D_i^l(t)$ — скорость убывания пакетов по l-му каналу, которая рассчитывается следующим образом:

$$D_i^l(t) = \begin{cases} A_i^l(t), & q_l(t) = 0, \\ \dfrac{A_i^l(t-d)}{\sum\limits_{j \in N_l} A_j^l(t-d)} C_l, & q_l(t) > 0, \end{cases}$$

где d — задержка, обусловливаемая временем нахождения пакетов в очереди, испытываемая трафиком, выходящим из l в момент времени t. Задержка d есть решение уравнения

$$\frac{q_l(t-d)}{C_l} = d.$$

К недостаткам классической ЖМ [4] и ее последующим известным модификациям (см., например, [5], авторы которой предложили способ интеграции ЖМ трафика и сетевого симулятора *ns*-2) следует отнести отсутствие возможности решить задачу учета рассогласованного характера действий пользователя. Отмеченный недостаток в известной мере был преодолен авторами работы [1]. В этой работе была предложена гибридная жидкостная модель (ГЖМ) информационных потоков в высокоскоростных КС, представляющая собой комбинацию ЖМ [4] и статистического варианта модели абстрактных источников трафика. Такая модель позволяет описывать информационные потоки в мультисервисных КС с учетом присущего протоколу *TCP* механизма обратного влияния загрузки сети на режим работы источника трафика, а также учитывать современные политики управления скоростью доступа к сети Интернет отдельных пользователей.

Обсуждаемая ГЖМ была доведена ее авторами до законченной последовательной (работающей на *CPU*) программной реализации [1]. Ее работоспособность подтверждена результатами сравнения вычисленных и полученных в работе [6] характеристик информационных потоков. В то же время опыт практического использования последовательной реализации ГЖМ показал, что временные затраты, необходимые для проведения расчетов с относительно небольшим числом информационных потоков, оказываются весьма значительными и возрастают одновременно с увеличением числа моделируемых потоков. В результате последовательная программная реализация ГЖМ не позволяла проводить моделирование высокоскоростных КС со сложной топологией вследствие крайне низкой скорости расчетов, что было продемонстрировано в работе [7].

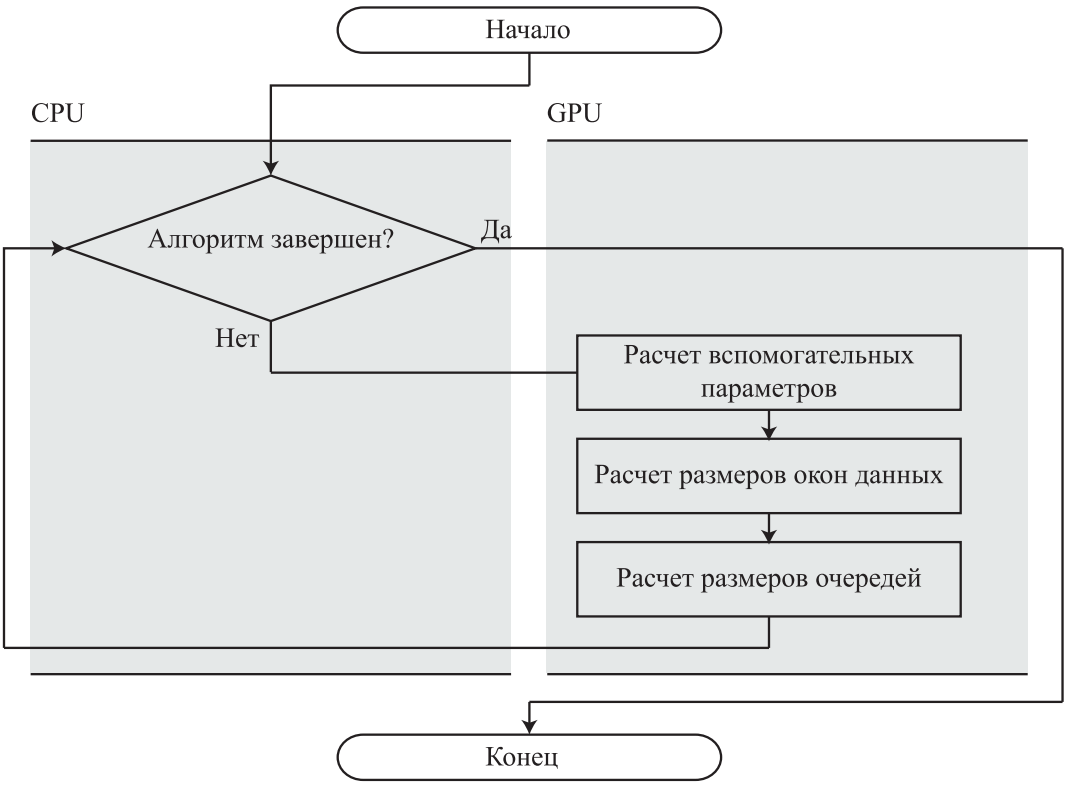

Рис. 1. Блок-схема алгоритма работы параллельной ГЖМ

В настоящее время один из наиболее эффективных подходов для повышения скорости вычислений заключается в использовании параллельных вычислений, реализованных с помощью технологии неспециализированных вычислений на графических процессорах (*General-purpose computing for graphics processing units*, *GPGPU*) [8]. В работе [9] была выдвинута, обоснована, а в работе [10] подтверждена гипотеза о целесообразности разработки на основе данной технологии параллельных программных реализаций как классической ЖМ, так и ГЖМ. На рис. 1 представлена блок-схема параллельной ГЖМ.

В настоящей статье обсуждаются результаты сравнительного анализа свойств синтетического интернет-трафика, сгенерированного при использовании последовательной и параллельной программных реализаций ГЖМ для КС со сложной топологией. В работе представлены полученные в ходе исследований оценки ускорения расчетов, обеспечиваемых использованием технологии *GPGPU*, при различном числе информационных потоков и маршрутизаторов КС.

Анализ результатов моделирования информационных потоков в КС со сложной топологией с помощью последовательной и параллельной программных реализаций ГЖМ

Для подтверждения работоспособности параллельной реализации ГЖМ было проведено два эксперимента. В ходе исследования для каждого из них были получены характеристики информационных потоков, передаваемых в КС со сложной топологией. Эксперименты проводили на последовательной и параллельной реализациях ГЖМ, после чего проводили сравнение полученных значений.

Для экспериментов использовали КС, состоящую из двух маршрутизаторов, на входы которых поступали информационные потоки, обусловленные запросами пользователей. Топология моделируемой сети представлена на рис. 2.

В выбранной топологии КС источники F генерировали информационные потоки, каждый из которых проходил через маршрутизаторы R_1 и R_2 до источника S. Размер каждой группы источников, подключенных к маршрутизаторам, составлял n единиц. Таким образом, общее число источников составляло $2n$. При этом время распростране-

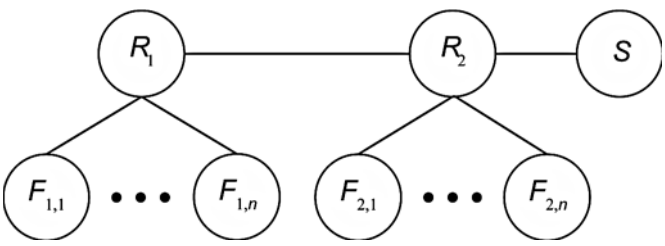

Рис. 2. Базовая топология КС, использованная в проведенных экспериментах:

R — маршрутизаторы; F — источники трафика; S — приемник трафика

ния сигналов по всем каналам связи варьировалось в зависимости от проводимого эксперимента. Время распространения сигнала в канале связи, соединяющем маршрутизатор R_2 и приемник S, принималось равным нулю.

Эксперимент 1. Цель данного эксперимента заключалась в оценке соответствия входных и выходных нагрузок маршрутизаторов. Характеристики формируемых нагрузок обусловливаются параметрами алгоритма предоставления гарантированной скорости доступа и характеристиками активности источников трафика. В рамках данного эксперимента общее число моделируемых источников трафика составляло $N = 512$ ($n = 256$). Скорость передачи данных каждого источника обусловливалась параметрами алгоритма предоставления гарантированной скорости доступа *Rate Limit* (размеры согласованного и расширенного всплесков, гарантируемая скорость доступа). Скорость работы маршрутизаторов при этом не ограничивалась. В проведенных экспериментах были использованы следующие значения параметров ГЖМ:

- время моделирования — 50 с;
- время двойного оборота (*RTT*) — 50...70 мс;
- согласованный размер всплеска алгоритма *Rate Limit* — 0 Кбит;
- расширенный размер всплеска алгоритма *Rate Limit* — 50 Кбит;
- гарантируемая скорость доступа алгоритма *Rate Limit* — 1,5 Мбит/с;
- длительность активности пользователей — 0,3...1 с;
- среднее время между запросами — 0,05 с.

Результаты расчетов для параллельной и последовательной программных реализаций представлены на рис. 3 и 4 соответственно.

Для сравнения результатов расчетов были использованы следующие количественные характеристики информационных потоков в КС: минимальная скорость передачи данных, средняя скорость передачи данных, максимальная скорость передачи данных для входных и выходных нагрузок маршрутизаторов. Выбранные характеристики представлены в табл. 1.

Из данных табл. 1 видно, что значения выбранных параметров согласуются друг с другом. Имеющиеся незначительные различия в результатах расчетов объясняются случайным характером параметров ГЖМ (время активности и простоя "мулов", время двойного оборота сигнала).

Эксперимент 2. Цель данного эксперимента заключалась в оценке соответствия входных и выходных нагрузок маршрутизаторов, характеристики которых определяются параметрами алгоритма активного управления очередью *RED* (нижний и верхний пороги срабатывания функции сброса пакетов и максимальная вероятность сброса пакетов) и характеристиками активности пользователей КС.

В качестве источников трафика были использованы следующие классы пользователей: "слоны" (пользователи, осуществляющие скачивание и передачу больших объемов данных (сотни Мбайт), сохраняющие активность на протяжении всего процесса моделирования); "мулы" — (пользователи, осуществляющие скачивание и передачу объемов данных среднего размера (сотни и тысячи Кбайт). В ходе моделирования длительности периодов активности

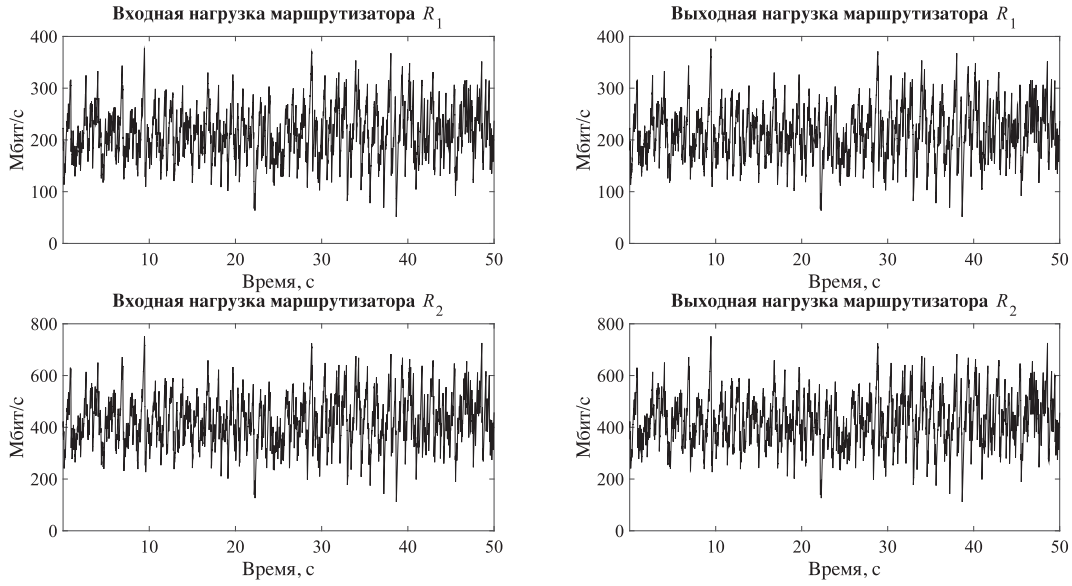

Рис. 3. Эксперимент 1. Параллельная программная реализация ГЖМ: зависимости "мгновенных" значений скорости передачи данных на входах и выходах маршрутизаторов от времени

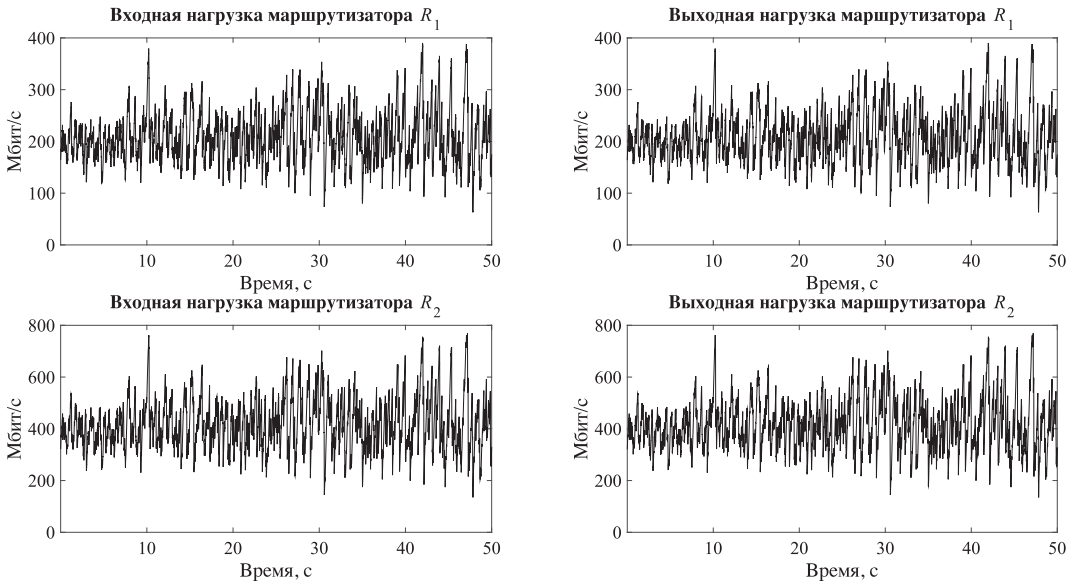

Рис. 4. Эксперимент 1. Последовательная программная реализация ГЖМ: зависимости "мгновенных" значений скорости передачи данных на входах и выходах маршрутизаторов от времени

Таблица 1

Сравнение скорости передачи данных для эксперимента 1

Параметры сравнения	Минимальная скорость, Мбит/с	Средняя скорость, Мбит/с	Максимальная скорость, Мбит/с
Характеристики, полученные последовательной программной реализацией			
Входная нагрузка маршрутизатора R_1	63,0134	208,8759	390,3031
Выходная нагрузка маршрутизатора R_1	63,0134	208,8756	390,3031
Входная нагрузка маршрутизатора R_2	134,7500	417,7901	769,1149
Выходная нагрузка маршрутизатора R_2	134,7500	417,7896	769,1149
Характеристики, полученные параллельной программной реализацией			
Входная нагрузка маршрутизатора R_1	51,3785	212,9495	376,4153
Выходная нагрузка маршрутизатора R_1	51,3785	212,9494	376,4153
Входная нагрузка маршрутизатора R_2	111,6676	423,4836	751,9642
Выходная нагрузка маршрутизатора R_2	111,6676	423,4829	751,9642

"мулов" (*ON*) варьировались в диапазоне 5...30 с, среднее время между запросами "мулов" составляло 3 с (период *OFF*). Отметим, что класс "мыши", к которому относятся пользователи, осуществляющие скачивание и передачу небольших объемов данных (десятки Кбайт), в экспериментах не учитывался.

Для эксперимента было выбрано следующее соотношение различных типов пользователей — 28 "слонов" и 100 "мулов". Каждый из выбранных типов пользователей был разделен между маршрутизаторами R_1 и R_2 на две одинаковые группы.

Расчеты проводили при следующих значениях параметров ГЖМ:

- время моделирования — 50 с;
- время двойного оборота сигнала (*RTT*) — 20...100 мс;
- размер буфера сетевого маршрутизатора — 1000 Кбайт;
- пропускная способность маршрутизатора R_1 — 25 Мбит/с;
- пропускная способность маршрутизатора R_2 — 50 Мбит/с;
- нижний порог алгоритма *RED* — 1 Кбит;
- верхний порог алгоритма *RED* — 250 Кбит;
- максимальная вероятность сброса пакетов для алгоритма *RED* — 0,2.

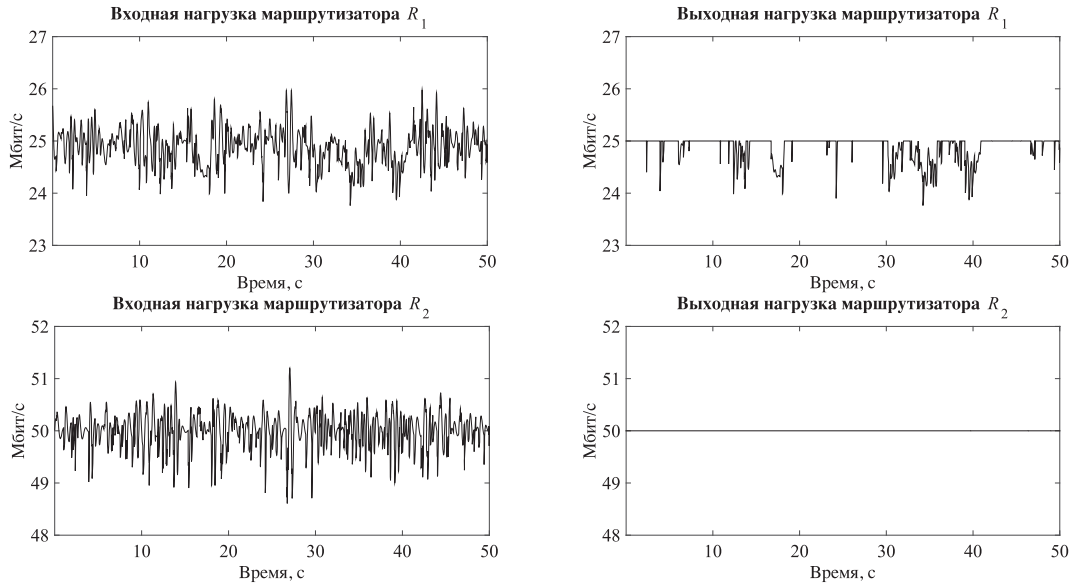

Рис. 5. Эксперимент 2. Параллельная программная реализация ГЖМ: зависимости "мгновенных" значений скорости передачи данных на входах и выходах маршрутизаторов от времени

Результаты расчетов, полученные с помощью параллельной и последовательной реализаций ГЖМ, представлены на рис. 5 и 6 соответственно.

Выбранные характеристики параметров моделируемых информационных потоков представлены в табл. 2.

Из данных табл. 2 видно, что как и в эксперименте 1 значения выбранных параметров согласуются друг с другом.

Таким образом, результаты, представленные в табл. 1 и 2, позволяют сделать обоснованный вывод о работоспособности разработанной параллельной программной реализации ГЖМ на основе технологии *GPGPU* для КС со сложными топологиями.

Сравнительный анализ скорости расчетов последовательной и параллельной реализаций гибридной жидкостной модели

После получения подтверждения работоспособности параллельной программной реализации ГЖМ на основе технологии *GPGPU* было проведено измерение времени вычислений, затрачиваемых последовательной и параллельной программными реали-

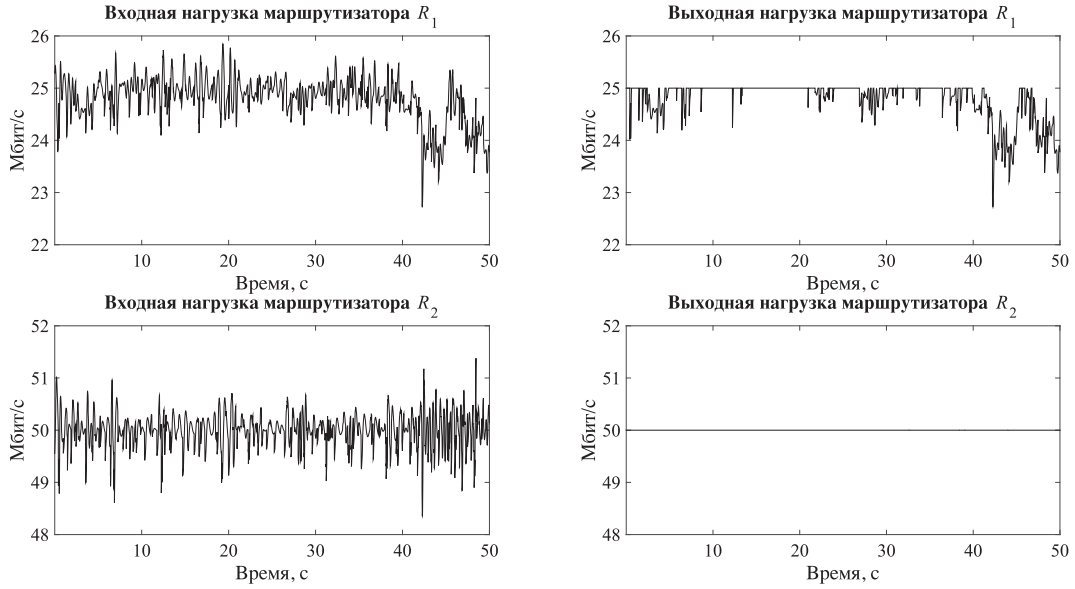

Рис. 6. Эксперимент 2. Последовательная программная реализация ГЖМ: зависимости "мгновенных" значений скорости передачи данных на входах и выходах маршрутизаторов от времени

Таблица 2

Сравнение скорости передачи данных для эксперимента 2

Параметры сравнения	Минимальная скорость, Мбит/с	Средняя скорость, Мбит/с	Максимальная скорость, Мбит/с
Характеристики, полученные последовательной программной реализацией			
Входная нагрузка маршрутизатора R_1	22,7283	24,8002	25,8484
Выходная нагрузка маршрутизатора R_1	22,7283	24,8021	25,0000
Входная нагрузка маршрутизатора R_2	48,3563	50,0024	51,3790
Выходная нагрузка маршрутизатора R_2	50,0000	50,0000	50,0000
Характеристики, полученные параллельной программной реализацией			
Входная нагрузка маршрутизатора R_1	23,7630	24,8905	25,9784
Выходная нагрузка маршрутизатора R_1	23,7630	24,8927	25,0000
Входная нагрузка маршрутизатора R_2	48,6080	50,0012	51,2028
Выходная нагрузка маршрутизатора R_2	50,0000	50,0000	50,0000

зациями ГЖМ при моделировании КС со сложной топологией.

Методика проведения экспериментов. Принципиальная схема топологий КС, использованных в проведенных экспериментах, представлена на рис. 7.

На рис. 7 видно, что в обсуждаемых экспериментах топология сети представляла собой группу небольших несвязанных между собой сетей. Каждая сеть состояла из двух *TCP*-источников, одного маршрутизатора, одного приемника и трех каналов связи. Общее число сегментов сети для k-го эксперимента составляло $m = 2^{k-1}$, $k = 1, ..., 10$.

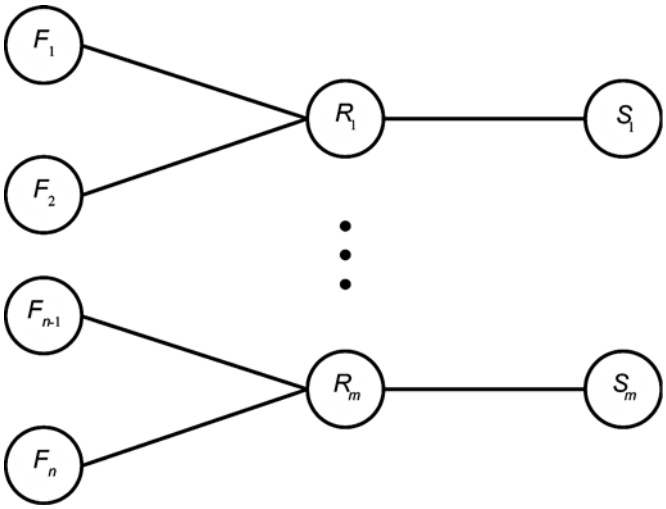

Рис. 7. Принципиальная схема топологии моделируемых КС для k-го эксперимента:

F — *TCP*-источник интернет-трафика; R —маршрутизатор; S — приемник нагрузки

Каждый *TCP*-источник генерировал нагрузку на протяжении всего времени моделирования для обеспечения максимальной нагрузки на вычислительные алгоритмы и аппаратные ресурсы. В каждом эксперименте варьировались следующие параметры ГЖМ: число моделируемых *TCP*-источников интернет-трафика; число маршрутизаторов; число приемников. Остальные параметры ГЖМ оставались неизменными. Данный подход позволил проводить идентичные эксперименты с помощью обеих программных реализаций ГЖМ.

В k-м эксперименте использовали следующие параметры ГЖМ:

- время моделирования — 1000 мс;
- число пользователей — $n = 2^k$;
- число маршрутизаторов — $m = 2^{k-1}$.

Замеры времени вычислений проводили перед началом моделирования и после его окончания без учета операций инициализации. Для каждого набора параметров, соответствующих эксперименту k, проводили 20 независимых замеров, которые впоследствия усреднялии Замеры проводили для последовательной и параллельной программных реализаций ГЖМ. После этого полученные значения сравнивали друг с другом.

Для исключения влияния внешних факторов на производительность каждой из программных реализаций в экспериментах с последовательной и параллельной программными реализациями ГЖМ использовали одинаковое аппаратное обеспечение и операционную систему.

Анализ результатов экспериментов. Сравнение усредненных по ансамблю независимых реализаций решений СДУ (1), (2), полученных с помощью последовательной и параллельной реализаций ГЖМ, показали, что их отличия друг от друга оказываются

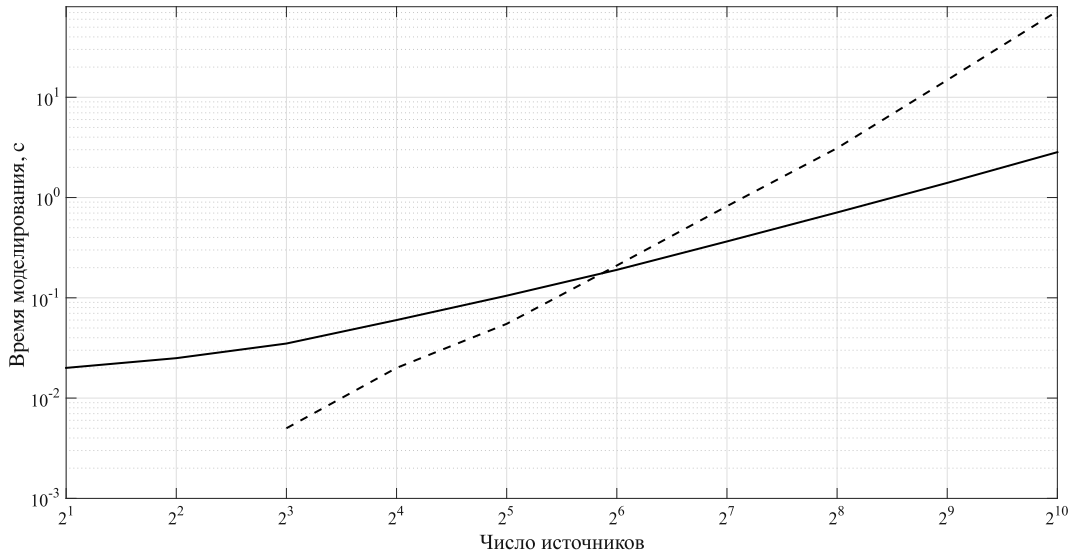

Рис. 8. Зависимость времени решения СДУ (1), (2) от числа *TCP*-источников: штриховая линия — последовательная реализация ГЖМ; сплошная линия — параллельная реализация ГЖМ

в пределах погрешности, обусловленной конечной точностью компьютерных вычислений.

Усредненные по ансамблю реализаций замеры времени моделирования для десяти вариантов экспериментов представлены на рис. 8.

На рис. 8 видно, что скорость вычислений параллельной реализации гибридной жидкостной модели при числе информационных потоков > 2^6 оказывается выше, чем соответствующий последовательный вариант обсуждаемой модели. Например, когда число пользователей составляет 2^{10}, скорость параллельной реализации ГЖМ оказывается в 25 раз быстрее последовательной реализации.

Для подтверждения данного вывода в табл. 3 приведены количественные значения времени вычислений.

В тоже время, если число пользователей относительно невелико (<2^6), скорость вычислений последовательной реализации ГЖМ, напротив, оказывается выше. Например, при моделировании одного сегмента КС, состоящего из двух источников и одного маршрутизатора, время расчета модели составляет приблизительно 0,02 с. Данный результат объясняется тем, что вне зависимости от количества пользователей в параллельной реализации ГЖМ изначально должны быть инициализированы вычислительные

Таблица 3

Числовые значения времени моделирования различного числа сегментов КС

Число источников	Время моделирования последовательной программной реализацией, с	Время моделирования параллельной программной реализацией, с	Ускорение
2^1	$99{,}9\cdot10^{-5}$	$20{,}0\cdot10^{-3}$	0,0499
2^2	$20{,}0\cdot10^{-4}$	$25{,}0\cdot10^{-3}$	0,0800
2^3	$50{,}0\cdot10^{-4}$	$35{,}0\cdot10^{-3}$	0,1429
2^4	$20{,}0\cdot10^{-3}$	$60{,}0\cdot10^{-3}$	0,3334
2^5	$55{,}0\cdot10^{-3}$	$10{,}5\cdot10^{-2}$	0,5238
2^6	$21{,}0\cdot10^{-2}$	$19{,}0\cdot10^{-2}$	1,1052
2^7	$82{,}0\cdot10^{-2}$	$36{,}5\cdot10^{-2}$	2,2466
2^8	$31{,}1\cdot10^{-1}$	$71{,}0\cdot10^{-2}$	4,3803
2^9	$14{,}9\cdot10^{0}$	$14{,}0\cdot10^{-1}$	10,6072
2^{10}	$72{,}2\cdot10^{0}$	$28{,}4\cdot10^{-1}$	25,4832

ядра графического процессора. По этой причине время, затрачиваемое на данную процедуру, определяет минимально возможное время, за которое модель может быть полностью просчитана. Практически линейная зависимость времени моделирования от количества пользователей при их небольшом числе свидетельствует о неполной загрузке *GPU*, большая часть вычислительных ресурсов которого простаивает.

Заключение

Проведено сравнение последовательной и параллельной программных реализаций ГЖМ, адаптированных для моделирования информационных потоков в КС со сложными топологиями.

Сравнение количественных характеристик информационных потоков на входе маршрутизатора (минимальной скорости передачи данных, средней скорости передачи данных, максимальной скорости передачи данных) позволяет сделать вывод о согласованности результатов, которые получены в ходе исследования представленных программных реализаций.

Исследованы зависимости скорости выполнения вычислений последовательной и параллельной программных реализаций. Полученные результаты показывают, что скорость вычислений параллельной программной реализации ГЖМ при большом числе пользователей ($>2^6$) оказывается выше, чем скорость вычислений последовательной программной реализации. При малом числе пользователей, наоборот, скорость работы однопроцессорного варианта программной реализации ГЖМ оказывается выше. Этот факт объясняется тем, что независимо от числа пользователей системе требуется время на инициализацию вычислительных ядер на графическом процессоре.

Список литературы

1. **Гребенкин М. К., Поршнев С. В.** Гибридная жидкостная модель магистрального интернет-канала. Saarbrucken: LAP LAMBERT Academic Publishing GmbH & Co. KG, 2012. 163 с.
2. **Гнеденко Б. В., Коваленко И. Н.** Введение в теорию массового обслуживания. М.: КомКнига, 2005. 400 с.
3. **Хинчин А. Я.** Математические методы теории массового обслуживания. Тр. Мат. Института им. В. А. Стеклова АН СССР. 1955. 122 с.
4. **Vishal Misra, Wei-Bo Gong, Don Towsley.** Stochastic Differential Equation Modeling and Analysis of TCP Windowsize Behavior// Proceedings of IFIP WG 7.3 Performance, November, 1999. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.45.9562&rep=rep1&type=pdf
5. **Marsan M. A., Garetto M., Giaccone P., Leonardi E., Schiattarella E., Tarello A.** Using Partial Differential Equations to Model TCP Mice and Elephants in Large IP Networks // INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies. 2004. Vol. 4. P. 2821—2832.
6. **Гребенкин М. К., Поршнев С. В.** Программная реализация гибридной жидкостной модели информационных потоков в высокоскоростных магистральных интернет-каналах. Свидетельство о государственной регистрации программ для ЭВМ № 2012616118. Заявка № 2012613786. Дата поступления 11 мая 2012 г. Дата регистрации в Реестре программ для ЭВМ 4 июля 2012 г.
7. **Басавин Д. А., Поршнев С. В.** Последовательная и параллельная реализации гибридной жидкостной модели информационных потоков в компьютерных сетях со сложной топологией // Cloud of Science. 2017. Т. 4. № 2. С. 216—223.
8. **Xu Zh., Bagrodia R.** GPU-Accelerated Evaluation Platform for High Fidelity Network Modeling // Proceeding PADS'07 Proceedings of the 21st International Workshop on Principles of Advanced and Distributed Simulation. 2007. P. 131—140.
9. **Басавин Д. А., Поршнев С. В.** О целесообразности использования графических процессоров при моделировании крупных телекоммуникационных систем // Актуальные вопросы в научной работе и образовательной деятельности: сборник научных трудов по материалам Международной научно-практической конференции 31 января 2013 г.: в 13 частях. Часть 2. М-во обр. и науки РФ. Тамбов: Бизнес-Наука-Общество. 2013. С. 12—15.
10. **Басавин Д. А., Поршнев С. В.** Параллельная гибридная жидкостная модель высокоскоростных информационных потоков в магистральных интернет-каналах // Естественные и технические науки. 2013. № 1. С. 317—326.

Comparison of Sequential and Parallel Software Implementations of the Hybrid Fluid Model of Information Flows for Computer Networks with Complex Topology

D. A. Basavin[1], basavind@gmail.com, **S. V. Porshnev**[2], sergey_porshnev@mail.ru,
D. A. Petrosov[1], scorpionss2002@mail.ru,

[1] Belgorod State Agricultural University named after V. Gorin, Mayskiy, Belgorod region, 308503, Russian Federation
[2] Ural Federal University named after the first President of Russia B. N. Yeltsin, Ekaterinburg, 620002, Russian Federation

Corresponding author:

Basavin Dmitry A., Assistant, Belgorod State Agricultural University named after V. Gorin, Mayskiy, Belgorod Region, 308503 Russian Federation
E-mail: basavind@gmail.com

Received on December 01, 2017
Accepted on December 07, 2017

The article discusses the analysis and comparison of software implementations of the hybrid fluid model (HFM) of information flows in computer networks with complex topologies. The sequential version implemented on the central processing unit (CPU) and the parallel version implemented on the graphics processing unit (GPU) using general-purpose computing for graphics processing units (GPGPU) technology are considered.

During the analysis of the GPU based HFM implementation, it was verified against the CPU based implementation on a series of experiments. The results of the experiments confirmed the convergence of the results obtained by both versions and are presented in the article. A number of quantitative information flows characteristics were compared graphically and numerically to verify the results. Thus characteristics were the minimum, average and maximum input and output aggregated flow rates on routers.

A number of experiments were performed for performance estimation and acceleration of the parallel HFM relatively to sequential. The parameters of the experiments were selected to maximize the use of available computing resources. The size of the simulated network for each experiment increased exponentially. For each experiment, measurements of computation time for sequential and parallel implementations of HFM were carried out, after which they compared graphically and numerically.

During the comparison of software implementations, the hypothesis was confirmed by the advisability of using GPGPU technology to accelerate the calculations of HFM. It clearly demonstrated that the parallel software implementation makes sense to apply to the number of information flows >2^6.

Keywords: *Internet traffic, computer networks, parallel hybrid fluid model, modeling, GPGPU*

For citation:

Basavin D. A., Porshnev S. V., Petrosov D. A. Comparison of Sequential and Parallel Software Implementations of the Hybrid Fluid Model of Information Flows for Computer Networks with Complex Topology, *Programmnaya Ingeneria,* 2018, vol. 9, no. 2, pp. 59—68.

DOI: 10.17587/prin.9.59-68

References

1. **Grebenkin M. K., Porshnev S. V.** *Gibridnaja zhidkostnaja model' magistral'nogo internet-kanala* (A hybrid fluid model is the backbone of the Internet channel), Saarbrucken, LAP LAMBERT Academic Publishing GmbH & Co. KG, 2012. 163 p. (in Russian).

2. **Gnedenko B. V., Kovalenko I. N.** *Vvedenie v teoriju massovogo obsluzhivanija* (Introduction to the theory of mass service), Moscow, KomKniga, 2005, 400 p. (in Russian).

3. **Hinchin A. Ja.** *Matematicheskie metody teorii massovogo obsluzhivanija* (Mathematical methods of theory of mass service), Tr. Mat. Instituta im. V. A. Steklova AN SSSR, 1955, 122 p. (in Russian).

4. **Vishal Misra, Wei-Bo Gong, Don Towsley.** Stochastic Differential Equation Modeling and Analysis of TCP Windowsize Behavior, *Proceedings of IFIP WG 7.3 Performance*, November, 1999, available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.45.9562&rep=rep1&type=pdf

5. **Marsan M. A., Garetto M., Giaccone P., Leonardi E., Schiattarella E., Tarello A.** Using Partial Differential Equations to Model TCP Mice and Elephants in Large IP Networks, *INFOCOM 2004. Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies*, 2004, vol. 4, pp. 2821—2832.

6. **Grebenkin M. K., Porshnev S. V.** Programmnaja realizacija gibridnoj zhidkostnoj modeli informacionnyh potokov v vysokoskorostnyh magistral'nyh internet-kanalah. Svidetel'stvo o gosudarstvennoj registracii programm dlja JeVM № 2012616118. Zajavka № 2012613786. Data postuplenija 11.03.2012. Data registracii v Reestre programm dlja JeVM 4.07.2012 (in Russian).

7. **Basavin D. A., Porshnev S. V.** Posledovatel'naja i parallel'naja realizacii gibridnoj zhidkostnoj modeli informacionnyh potokov v komp'juternyh setjah so slozhnoj topologiej (Serial and parallel implementations of hybrid fluid model of information flows in networks with complex topology), *Cloud of Science,* 2017, vol. 4, no. 2, pp. 216—223 (in Russian).

8. **Xu Zh., Bagrodia R.** GPU-Accelerated Evaluation Platform for High Fidelity Network Modeling, *Proceeding PADS '07 Proceedings of the 21st International Workshop on Principles of Advanced and Distributed Simulation*, 2007, pp. 131—140.

9. **Basavin D. A., Porshnev S. V.** O celesoobraznosti ispol'zovanija graficheskih processorov pri modelirovanii krupnyh telekommunikacionnyh system (The feasibility of using graphics processing units for simulation of large telecommunication systems), *Aktual'nye voprosy v nauchnoj rabote i obrazovatel'noj dejatel'nosti: sbornik nauchnyh trudov po materialam Mezhdunarodnoj nauchno-prakticheskoj konferencii,* 31 January 2013, 13 parts. Part 2. M-vo obr. i nauki RF, Tambov, Biznes-Nauka-Obshhestvo, 2013, pp. 12—15 (in Russian).

10. **Basavin D. A., Porshnev S. V.** Parallel'naja gibridnaja zhidkostnaja model' vysokoskorostnyh informacionnyh potokov v magistral'nyh internet-kanalah (The parallel hybrid fluid model of high-speed information streams in the backbone Internet channels), *Estestvennye i tehnicheskie nauki,* 2013, no. 1. pp. 317—326 (in Russian).

УДК 004.056.53 DOI: 10.17587/prin.9.69-75

П. В. Закалкин, канд. техн. наук, сотр., e-mail: ansmed82@mail.ru,
П. В. Мельников, канд. техн. наук, сотр., e-mail: ansmed82@mail.ru,
Академия Федеральной службы охраны Российской Федерации, г. Орел

Система анализа программного обеспечения на предмет отсутствия недекларированных возможностей

Рассмотрена система анализа программного обеспечения на предмет отсутствия недекларированных возможностей. Такая система предназначена для защиты информационных ресурсов рабочих станций и серверов, их компонентов, программ или данных от несанкционированной деятельности. Она может использоваться для анализа исходного кода программного обеспечения, в том числе при проведении сертификационных испытаний программного обеспечения, на отсутствие недекларированных возможностей. Существующие системы анализа программного обеспечения обладают рядом недостатков, которые не позволяют сопоставить результаты исследований с контрольным экземпляром (версией) одного и того же программного обеспечения. Это обстоятельство приводит к высоким временны́м затратам (за счет полного цикла проверок файлов исходных текстов различных версий программного обеспечения), а также к снижению доверия к результатам исследований.

Таким образом, возникает необходимость разработки системы анализа программного обеспечения на предмет отсутствия недекларированных возможностей, обеспечивающей повышение достоверности результатов анализа программного обеспечения путем фиксации исходного состояния, проверки избыточности исходных текстов и анализа контекстной информации в исходных текстах исследуемого программного обеспечения.

Ключевые слова: *программное обеспечение, недекларированные возможности, анализ исходного кода, сертификационные испытания*

Обеспечение безопасности информации, которая хранится и обрабатывается в различных информационных системах, подразумевает использование совокупности программно-технических и организационных средств и мер защиты. Важная роль при этом отводится средствам поиска дефектов (недостатков) программного кода, не декларированных спецификацией разработчика и не выявленных на этапе тестирования [1—3].

Причины возникновения дефектов кода различны и могут быть как случайными, так и преднамеренными. Дефекты кода не только негативно сказываются на ресурсоемкости и скорости выполнения программ, но и несут угрозу безопасности. Использование ряда дефектов программного кода, наличие которых не было своевременно обнаружено, потенциально способно предоставить злоумышленникам возможности по нарушению конфиденциальности, целостности и доступности информации. По данным исследования, проведенного по заказу Национального института стандартов и технологий США, убытки, возникающие вследствие недостаточно развитой инфраструктуры устранения уязвимостей и критических ошибок в программном обеспечении (ПО), составляют 22...60 млрд долларов США в год [4]. Указанные обстоятельства обусловливают необходимость совершенствования инструментария поиска и нейтрализации дефектов программного кода (ошибок программирования, уязвимостей, программных закладок и недекларированных возможностей).

В рамках настоящей статьи под недекларированными возможностями и программными закладками будем понимать приведенные далее.

Недекларированные возможности — функциональные возможности программного обеспечения, не описанные или не соответствующие описанным в документации, при использовании которых возможно нарушение конфиденциальности, доступности или целостности обрабатываемой информации [5]. Реализацией недекларированных возможностей, в частности, являются программные закладки.

Программные закладки — преднамеренно внесенные в программное обеспечение функциональные объекты, которые при определенных условиях

(входных данных) инициируют выполнение не описанных в документации функций программного обеспечения, приводящих к нарушению конфиденциальности, доступности или целостности обрабатываемой информации [5].

Известные системы контроля отсутствия недекларированных возможностей в программном обеспечении [6—8] имеют ряд недостатков, среди которых:

— отсутствие процедуры фиксации исходного состояния исследуемого программного обеспечения;

— отсутствие механизма проверки избыточности исходных текстов на уровне файлов;

— отсутствие механизма извлечения и анализа контекстной информации.

Перечисленные недостатки не позволяют сопоставить результаты исследований с контрольным экземпляром (версией) одного и того же ПО, что приводит к высоким временны́м затратам (за счет полного цикла проверок файлов исходных текстов различных версий ПО), а также к снижению доверия к результатам исследований.

Таким образом, возникает необходимость разработки системы анализа ПО на предмет отсутствия недекларированных возможностей, которая обеспечивает повышение достоверности результатов анализа ПО путем фиксации исходного состояния, проверки избыточности исходных текстов и анализа контекстной информации в исходных текстах исследуемого ПО.

Предлагаемая система анализа ПО на предмет отсутствия недекларированных возможностей приведена на рис. 1.

Система анализа программного обеспечения на предмет отсутствия недекларированных возможностей содержит следующие блоки: *1* — сохранения найденных потоков управления исследуемого ПО; *2* — хранения файлов исходных текстов ПО; *16* — фиксации исходного состояния исходных текстов ПО; *17* — проверки избыточности исходных текстов; *18* — анализа и систематизации контекстной информации; *3* — установки контрольных точек в файлы исходных текстов ПО; *4* — компиляции файлов исходных текстов ПО с установленными контрольными точками; *5* — выполнения скомпилированных файлов ПО; *6* — сохранения перечня пройденных контрольных точек; *7* — сохранения трасс маршрутов выполнения скомпилированных файлов ПО; *8* — хранения перечня потенциально опасных программных конструкций; *9* — поиска потенциально опасных программных конструкций в файлах исходных текстов ПО; *10* — экспертной корректировки результатов автоматизированного анализа; *15* — тестирования потенциально опасных программных конструкций; *11* — сохранения итоговых результатов.

Отличительной чертой предлагаемой системы является наличие блоков фиксации исходного состояния исходных текстов ПО, блока проверки избыточности исходных текстов и блока анализа и систематизации контекстной информации.

Проведение контроля исходного состояния исследуемого ПО, контроля полноты и отсутствия избыточности исходных текстов, а также анализа контекстной информации позволяет проводить сертификационные исследования ПО согласно п. 2 и пп. 3.1 перечня требований к уровням контроля, представленного в работе [5].

Повышение достоверности результатов исследования исходных текстов за счет предотвращения внесения в них несанкционированных изменений осуществляется фиксацией исходного состояния файлов исследуемого ПО. Блок фиксации исходного состояния предназначен для выполнения расчета контрольных сумм исходных текстов исследуемого ПО. Данный блок реализует функции расчета, сравнения и хранения контрольных сумм.

Схема блока фиксации исходного состояния исходных текстов ПО представлена на рис. 2.

Данный блок состоит из трех элементов: блок 16.1, в котором проводится расчет контрольных сумм файлов исходных текстов, блок 16.2, в котором осуществляется сравнение полученных контрольных сумм с эталонными (указанными в документации, поставляемой с исследуемым ПО), и блок 16.3, в котором осуществляется сохранение результатов контрольного суммирования в файл отчета. Реализация блоков 16.1 и 16.2, входящих в состав блока 16, возможна на основе сертифицированных программных средств. К числу таких средств относятся: "Средство фиксации и контроля исходного состояния

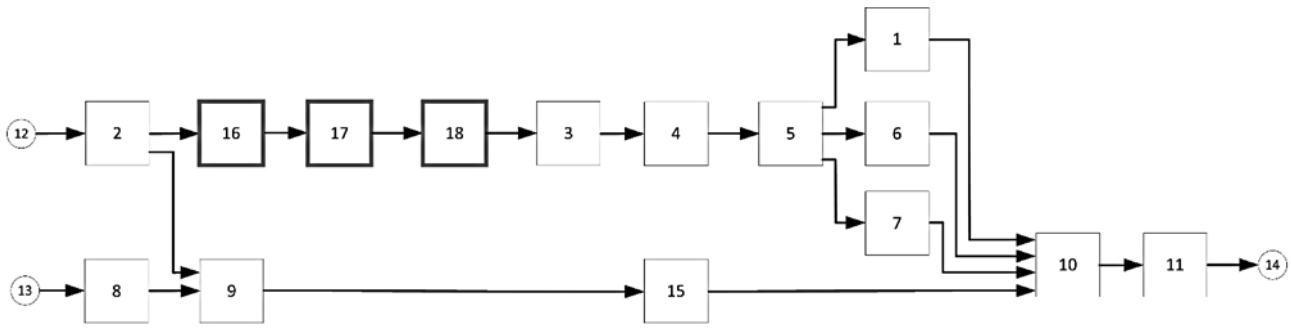

Рис. 1. Система анализа программного обеспечения на предмет отсутствия недекларированных возможностей

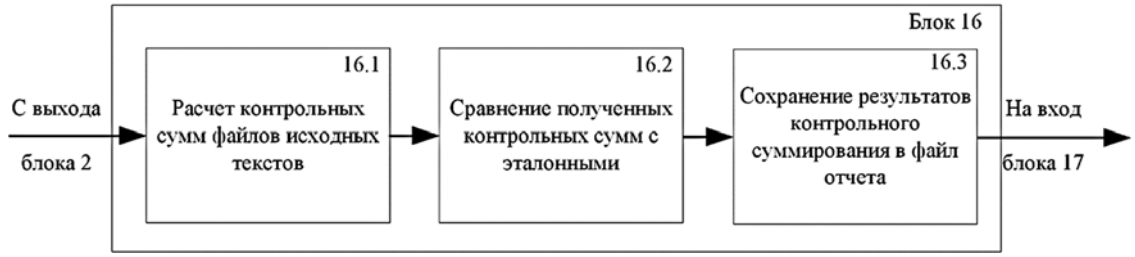

Рис. 2. Схема блока фиксации исходного состояния исходных текстов программного обеспечения

программного комплекса ФИКС" или "ПИК—Эшелон". Однако не все алгоритмы контрольного суммирования являются стойкими к коллизиям. В связи с этим обстоятельством в качестве алгоритма контрольного суммирования в рамках рассматриваемой системы предлагается использовать действующий Российский криптографический стандарт — ГОСТ Р 34.11—2012. Кроме этого, дополнительно к проверке контрольных сумм предлагается проводить сравнение файлов исходных текстов исследуемого ПО по размеру, что позволит различать файлы исходных текстов в случае получения одинаковых контрольных сумм. Блок-схема функционирования блока фиксации исходного состояния исходных текстов ПО представлена на рис. 3.

Проверка избыточности исходных текстов исследуемого ПО на уровне файлов осуществляется с помощью сборки специализированной версии ПО, содержащей предварительно внедренные идентификационные данные. Эти идентификационные данные сформированы таким образом, чтобы при обработке исполняемых файлов для каждого из них можно было получить полный и безызбыточный перечень использованных при его сборке файлов исходных текстов (из состава предоставленных на исследования). Использование данного блока повышает достоверность результатов исследования исходных текстов за счет устранения неиспользуемых файлов исходных текстов.

Схема блока 17 проверки избыточности исходных текстов представлена на рис. 4.

Данный блок состоит из перечисленных далее трех элементов. В блоке 17.1 в составе блока 17 осуществляется поиск в файлах исходных текстов

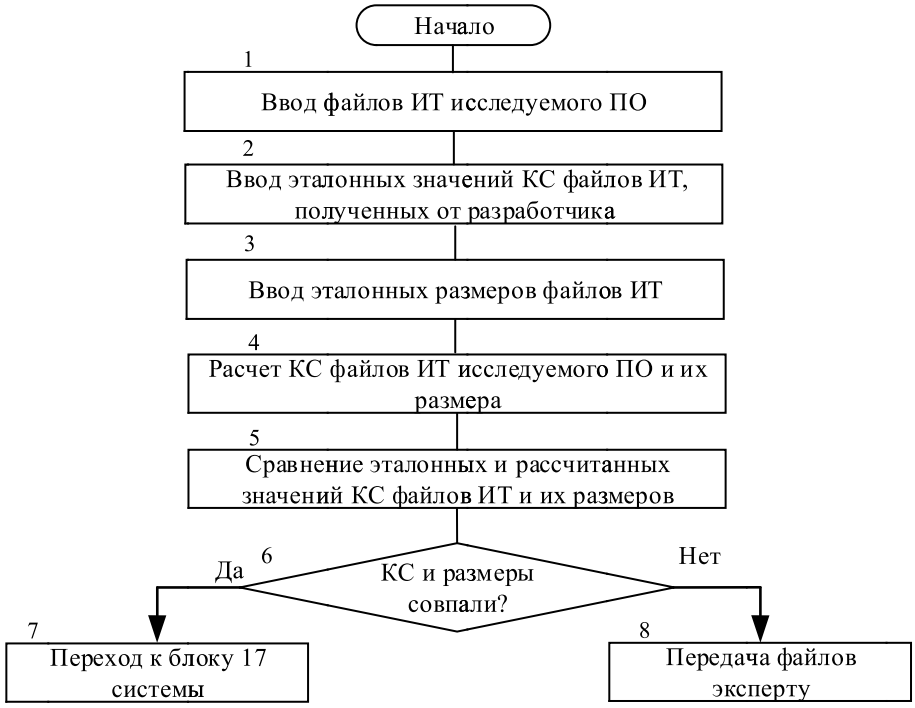

Рис. 3. Блок-схема функционирования блока фиксации исходного состояния исходных текстов программного обеспечения:

ИТ — исходные тексты; КС — контрольная сумма

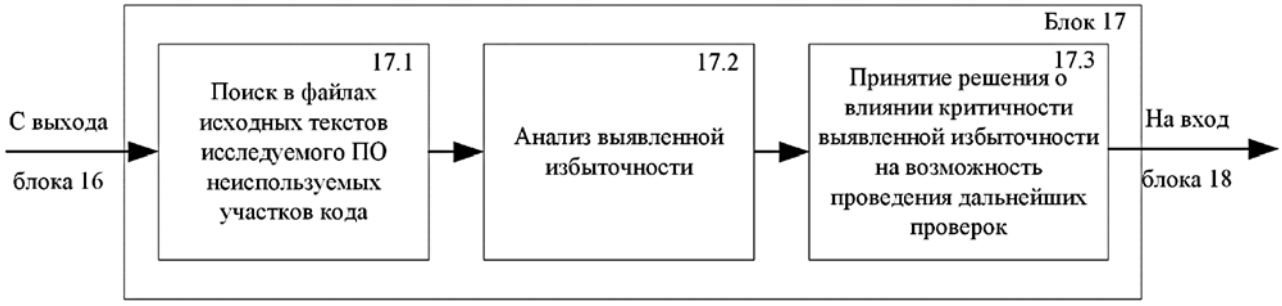

Рис. 4. Схема блока проверки избыточности исходных текстов

исследуемого ПО неиспользуемых участков кода. В блоке 17.2 проводится анализ выявленной избыточности, а в блоке 17.3 принимается решение о влиянии критичности выявленной избыточности на возможность и целесообразность проведения дальнейших проверок. Реализация блока 17.1 в плане поиска неиспользуемых участков исходных текстов возможна с использованием различного инструментария, в том числе и с использованием средств фиксации факта доступа к файлам. В операционной системе Linux, например, для этого используется подход, основанный на проверке параметра *"access time"* (данный параметр показывает время последнего доступа к файлу) соответствующего файла. В операционной системе Windows аналогичные функции выполняет Process Monitor. Данная программа в режиме реального времени отображает активность файловой системы, реестра, а также процессов и потоков. После реализации функций блоком 17.1, выявленная избыточность передается на вход блока 17.2. В данном блоке оператором осуществляется ее экспертный анализ, на основе результатов которого принимается решение в блоке 17.3.

Данный блок предназначен для выделения из состава исходных текстов только тех файлов, которые используются при сборке исследуемого проекта. Для операционных систем семейства Windows данный блок реализован посредством системного хука на функцию *CreateFile* библиотеки kernel32.dl и анализа ее параметров. В операционной системе Microsoft Windows хуком (*hook*) называют механизм перехвата событий с использованием особой функции (таких как передача сообщений Windows, ввод с мыши или клавиатуры) до того, как они дойдут до приложения.

Для операционной системы семейства Linux используется файловый параметр *atime*, который изменяет свое значение при всех попытках доступа к файлам (в опциях монтирования соответствующей файловой системы должна быть указана опция *strictatime*, иначе значение *atime* меняться не будет). Это позволяет отследить операции открытия файлов из директории исследуемых исходных текстов во время выполнения процедуры компиляции, а также сформировать безызбыточный перечень файлов исходных текстов. Файлы, к которым во время выполнения процедуры компиляции обращения не было, потенциально могут исключаться из проекта и в дальнейшем при проведении исследований могут не использоваться. А это, в свою очередь, влечет снижение трудозатрат на проведение исследований за счет возможного уменьшения объема исследуемых исходных текстов.

Блок-схема функционирования блока проверки избыточности исходных текстов представлена на рис. 5.

Блок извлечения и анализа контекстной информации предназначен для проведения исследования информации, содержащейся в комментариях исходных текстов. Данный блок реализует функции поиска и хранения комментариев к исходным текстам, а также функции для выделения фрагментов текста в коде, которые могут указывать на описание деструктивной или недокументированной функциональности. Использование данного блока повышает достоверность результатов исследования исходных текстов за счет проведения исследования описательной части в исходных текстах ПО.

Схема блока анализа и систематизации контекстной информации представлена на рис. 6.

Данный блок состоит из трех элементов: в блоке 18.1 осуществляется поиск контекстной информации в исследуемых файлах исходных текстов; в блоке 18.2 проводится анализ и систематизация выявленной контекстной информации; в блоке 18.3 делается вывод о наличии в контекстной информации указаний на потенциально опасные участки кода. Поиск контекстной информации в исследуемых файлах исходных текстов (блок 18.1) выполняется с помощью специализированных для конкретной ОС инструментальных механизмов (скриптов), реализующих функции поиска комментариев в файлах исходных текстов, а также путем выделения фрагментов текста в коде, которые могут указывать на описание негативных функций. После выполнения перечисленных действий выявленные комментарии передаются на вход блока 18.2, в котором оператором осуществляется их экспертный анализ, на основе результатов которого принимается решение в блоке 18.3.

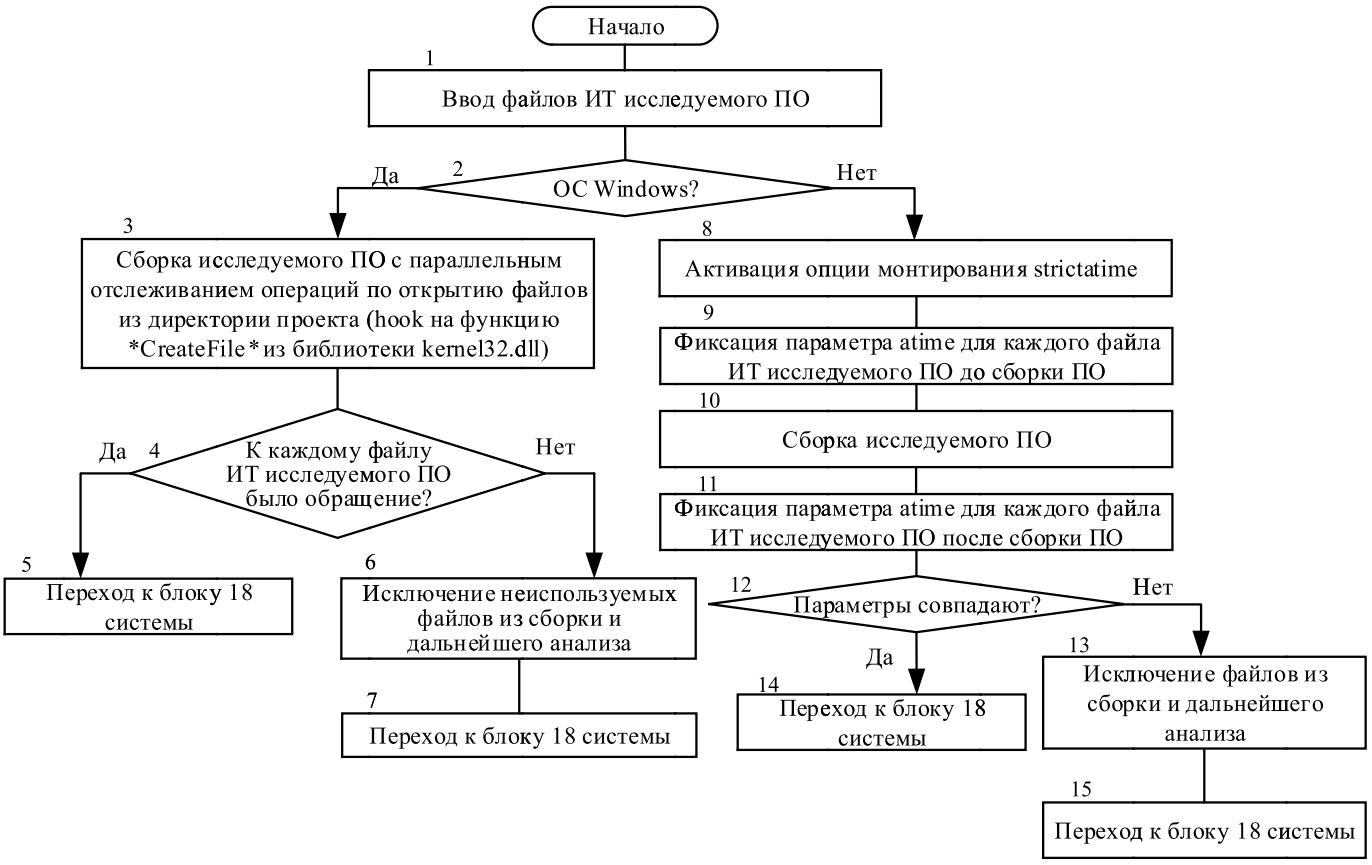

Рис. 5. Блок-схема блока проверки избыточности исходных текстов

Данный блок используется для выявления в исходном тексте исследуемого ПО сигнатур, указывающих на возможное наличие потенциально деструктивного функционала. Перечень используемых сигнатур (например, *crash*, *hook*, *password*, *lock*, *deny* и т. д.) хранится в базе данных. При анализе исходных текстов в данном блоке проводится извлечение комментариев и сравнение их с сигнатурами из базы данных. Отчет по результатам работы направляется эксперту для более тщательной проверки соответствующих участков кода.

Блок-схема функционирования блока анализа и систематизации контекстной информации представлена на рис. 7.

Таким образом, за счет введения блоков фиксации исходного состояния, проверки избыточности и анализа контекстной информации в исходных текстах исследуемого программного обеспечения повышается достоверность результатов анализа программного обеспечения на отсутствие недекларированных возможностей.

Предлагаемая система может быть использована при проведении сертификационных испытаний про-

Рис. 6. Схема блока анализа и систематизации контекстной информации

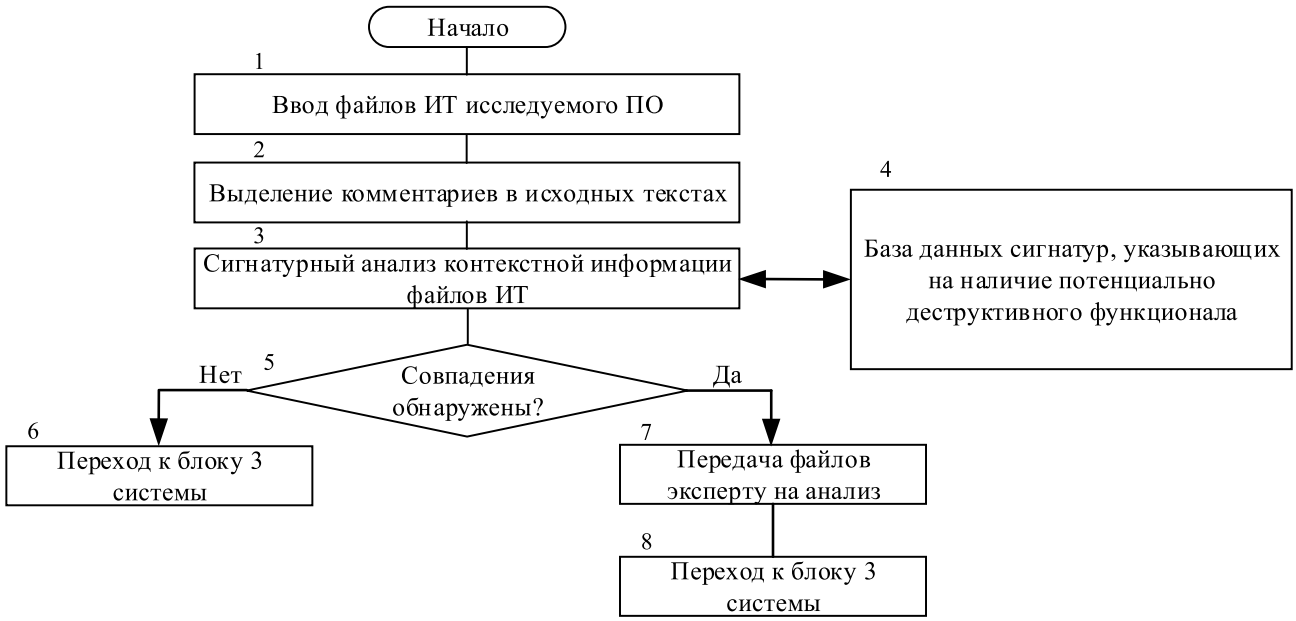

Рис. 7. Блок-схема блока анализа и систематизации контекстной информации

граммного обеспечения на отсутствие недекларированных возможностей.

Новизна способа подтверждается полученным патентом РФ № 2622622 "Система анализа программного обеспечения на отсутствие недекларированных возможностей" [9].

Список литературы

1. **Мельников П. В., Анисимов Д. В.** Проверка гарантий архитектуры программного обеспечения в процессе сертификационных испытаний (принципы построения диспетчера доступа) // Информационные системы и технологии. 2016. № 4. С. 112—120.

2. **Анисимов Д. В., Мельников П. В.** Проведение сертификационных исследований программного обеспечения с использованием технологии LLVM // Информационные системы и технологии. 2016. № 2. С. 99—104.

3. **Мельников П. В., Горюнов М. Н., Анисимов Д. В.** Подход к проведению динамического анализа исходных текстов программ // Вопросы кибербезопасности. 2016. № 3 (16) Спецвыпуск. С. 33—39.

4. **Контроль** уязвимостей в программных приложениях. URL: https://habrahabr.ru/company/jetinfosystems/blog/241353/

5. **Руководящий** документ "Защита от несанкционированного доступа к информации. Ч. 1. Программное обеспечение средств защиты информации. Классификация по уровню контроля отсутствия недекларированных возможностей". Утвержден решением председателя Государственной технической комиссии при Президенте Российской Федерации 04.06.1999 г.

6. **Пат. 2434272** Российская Федерация, МПК G06F 17/00. Система контроля отсутствия недекларированных возможностей в программном обеспечении / В. А. Минаков, В. В. Мирошников, В. В. Котрахов; заявитель и патентообладатель Федеральное государственное учреждение "Государственный научно-исследовательский испытательный институт проблем технической защиты информации Федеральной службы по техническому и экспортному контролю". 20100122971/08; заявл. 04.06.2010; опубл. 20.11.2011. 13 с.

7. **Пат. 2419135** Российская Федерация, МПК G06F 12/16, G06F 11/30. Система контроля отсутствия недекларированных возможностей в программном обеспечении / А. А. Бурушкин, К. П. Грищенко, В. А. Минаков, В. В. Мирошников; заявитель и патентообладатель Федеральное государственное учреждение "Государственный научно-исследовательский испытательный институт проблем технической защиты информации Федеральной службы по техническому и экспортному контролю". 2009136852/08; заявл. 05.10.2009; опубл. 20.05.2011. 11 с.

8. **Пат. 2434265** Российская Федерация, МПК G06F 11/00. Система контроля отсутствия недекларированных возможностей в программном обеспечении / А. А. Бурушкин, В. А. Минаков, В. В. Мирошников; заявитель и патентообладатель Федеральное государственное учреждение "Государственный научно-исследовательский испытательный институт проблем технической защиты информации Федеральной службы по техническому и экспортному контролю". 2010129013/08; заявл. 13.07.2010; опубл. 20.11.2011. 12 с.

9. **Пат. 2622622** Российская Федерация, МПК G06F 21/00, G06F 21/50, G06F 11/30, G06F 12/16. Система анализа программного обеспечения на отсутствие недекларированных возможностей / П. В. Закалкин, П. В. Мельников, М. Н. Горюнов, С. А. Воробьев, Д. В. Анисимов, К. Е. Петров; заявитель и патентообладатель Федеральное государственное казенное военное образовательное учреждение высшего образования "Академия Федеральной службы охраны Российской Федерации". 2016110533; заявл. 22.03.2016; опубл. 16.06.2017. 14 с.

System of the Analysis of the Software Regarding Lack of Undeclared Features

P. V. Zakalkin, ansmed82@mail.ru, **P. V. Mel'nikov,** ansmed82@mail.ru, Academy of the Federal Guard Service of the Russian Federation, Orel, 302034, Russian Federation

Corresponding author:

Zakalkin Pavel V., Researcher, Academy of the Federal Guard Service of the Russian Federation, Orel, 302034, Russian Federation
E-mail: ansmed82@mail.ru

Received on November 18, 2017
Accepted on December 07, 2017

In article the system of the analysis of the software regarding lack of not declared opportunities is considered. Such system is intended for protection of information resources of workstations and servers, their components, programs or data against unauthorized activity. It can be used for the analysis of a source code of the software, including, for example when carrying out certified tests of the software for lack of not declared opportunities. The existing systems of the analysis of the software possess a number of shortcomings which do not allow to compare results of researches with a control copy (version) of the same software. This circumstance leads to high time expenditure (at the expense of a full cycle of verifications of files of source texts of various versions of the software), and also to decrease in trust to results of researches.

Thus, there is a need to develop the system of the analysis of the software regarding lack of not declared opportunities, the reliability of results of the analysis of the software providing increase by fixing of an initial state, check of redundancy of source texts and the analysis of contextual information in source texts of the studied software.

Keywords: *software, not declared opportunities, analysis of a source code, certified tests*

For citation:

Zakalkin P. V., Mel'nikov P. V. System of the Analysis of the Software Regarding Lack of Undeclared Features, *Programmnaya Ingeneria,* 2018, vol. 9, no. 2, pp. 69—75.

DOI: 10.17587/prin.9.69-75

References

1. **Mel'nikov P. V., Anisimov D. V.** Proverka garantij arhitektury programmnogo obespechenija v processe sertifikacionnyh ispytanij (principy postroenija dispetchera dostupa) (Check of guarantees of architecture of the software in the course of certified tests (the principles of creation of the dispatcher of access)), *Informacionnye sistemy i tehnologii,* 2016, no. 4, pp. 112—120 (in Russian).

2. **Anisimov D. V., Mel'nikov P. V.** Provedenie sertifikacionnyh issledovanij programmnogo obespechenija s ispol'zovaniem tehnologii LLVM (Carrying out certified researches of the software with use of the LLVM technology), *Informacionnye sistemy i tehnologii*, 2016, no. 2, pp. 99—104 (in Russian).

3. **Mel'nikov P. V., Anisimov D. V., Gorjunov M. N.** Podhod k provedeniju dinamicheskogo analiza ishodnyh tekstov programm (Approach to carrying out the dynamic analysis of source texts programs), *Voprosy kiberbezopasnosti,* 2016, no. 3 (16), pp. 33—39 (in Russian).

4. **Kontrol'** ujazvimostej v programmnyh prilozhenijah (Control of vulnerabilities in software application), available at: https://habrahabr.ru/company/jetinfosystems/blog/241353/ (in Russian).

5. **The leading** document "Protection against unauthorized access to information. P.1. Software of means of information protection. Classification by the level of control of lack of not declared opportunities" (Zashhita ot nesankcionirovannogo dostupa k informacii. Ch. 1. Programmnoe obespechenie sredstv zashhity informacii. Klassifikacija po urovnju kontrolja otsutstvija nedeklarirovannyh vozmozhnostej). It is approved as the decision of the chairman of the State technical commission at the President of the Russian Federation 04.06.1999. (in Russian).

6. **Minakov V. A., Miroshnikov V. V., Kotrahov V. V.** Pat. 2434272 Russian Federation, MPK G06F 11/00. The monitoring system of lack of not declared opportunities in Software (Sistema kontrolja otsutstvija nedeklarirovannyh vozmozhnostej v programmnom obespechenii). Applicant and patent holder "State Research Test Institute of Problems of Technical Information Security of Federal Service on Technical and Export Control" Federal state institution. 20100122971/08; 04.06.2010 is declared; 20.11.2011 is published. 13 p.

7. **Burushkin A. A., Grishhenko K. P., Minakov V. A., Miroshnikov V. V.** Pat. 2419135 Russian Federation, MPK G06F 12/16, G06F 11/30. The monitoring system of lack of not declared opportunities in Software (Sistema kontrolja otsutstvija nedeklarirovannyh vozmozhnostej v programmnom obespechenii). Applicant and patent holder "State Research Test Institute of Problems of Technical Information Security of Federal Service on Technical and Export Control" Federal state institution. 2009136852/08; 05.10.2009 is declared; 20.05.2011 is published. 11 p.

8. **Burushkin A. A., Minakov V. A., Miroshnikov V. V.** Pat. 2434265 Russian Federation, MPK G06F 11/00. The monitoring system of lack of not declared opportunities in Software (Sistema kontrolja otsutstvija nedeklarirovannyh vozmozhnostej v programmnom obespechenii). Applicant and patent holder "State Research Test Institute of Problems of Technical Information Security of Federal Service on Technical and Export Control" Federal state institution. 2010129013/08; 13.07.2010 is declared; 20.11.2011 is published. 12 p.

9. **Zakalkin P. V., Mel'nikov P. V., Gorjunov M. N., Vorob'ev S. A., Anisimov D. V., Petrov K. E.** Pat. 2622622 Russian Federation, MPK G06F 21/00, G06F 21/50, G06F 11/30, G06F 12/16. System of the analysis of the software on absence not declared opportunities (Sistema analiza programmnogo obespechenija na otsutstvie nedeklarirovannyh vozmozhnostej). Applicant and patent holder Federal public state military educational institution of the higher education "Academy of Federal Guard Service of the Russian Federation". 2016110533; 22.03.2016 is declared; 16.06.2017 is published. 14 p.

УДК 004.04 DOI: 10.17587/prin.9.76-81

А. Бернадотт, канд. мед. наук, науч. сотр., e-mail: alexandra.bernadotte@gmail.com, Механико-математический факультет МГУ им. Ломоносова, Division of Physiological Chemistry I, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden

Анализ научного текста и новые мировые тенденции

Представлены результаты исследований, полученных на основе анализа публикаций журнала "Science" с 1996 по 2017 г., которые демонстрируют изменение фокуса научного знания, а именно — смещение акцента с теоретического и фундаментального знания в прикладную область науки. Кроме того, показаны тенденции к усилению коммерциализации в науке.

Анализ текста публикаций осуществлен с помощью классической модели определения близости тестов в их векторном представлении. В статье введена также модель алгоритмов кластеризации текстовых документов на основе предварительного формирования классов слов заранее выбранной, определенной направленности.

Ключевые слова: *анализ текста, семантический анализ, tf-idf, научный текст, познание, научные тенденции, кластерный анализ текста*

Научная гипотеза всегда выходит за пределы фактов, послуживших основой для ее построения.

В. И. Вернадский

Введение

В 2015 г. Джон Хорган опубликовал книгу под названием "Конец науки" (*The End of Science* [2]), в которой были представлены его размышления о смерти науки, о тенденциях, свидетельствующих о значительном изменении качества научного знания. Действительно, в настоящее время многие ученые, ретроспективно отслеживая предмет своих исследований, интуитивно замечают небольшие, а иногда даже довольно резкие изменения, произошедшие в научном мире с конца 1980-х годов. Изменения эти затрагивают все составляющие науки, включая семантику, методологию, аксиоматические основы и даже научные цели.

Исходя из предположения, что научный текст отражает тенденции в научном мире, имеет смысл проанализировать публикации в научных журналах, соответствующие предполагаемому периоду трансформации в науке. В настоящей работе представлены результаты анализа полнотекстовых публикаций журнала "Science" за период с 1996 по 2017 г.

Прежде чем представить методы, положенные в основу такого анализа, и его результаты, опишем интуитивно воспринимаемые тенденции в науке, к которым приводит беглый, неформальный анализ публикуемых результатов исследований.

Нарушенное равновесие теории и практики. Исходя из неформализованного наблюдения за большим объемом опубликованных научных статей, можно выделить несколько тенденций. Одной из текущих научных тенденций является существенный сдвиг от теории к практике с акцентом на эмпирическую составляющую в исследованиях. "Пустое теоретизирование" в таком подходе к исследованиям со временем действительно может оказаться пустым множеством. Теоретические подходы, по мнению некоторых современных исследователей, уходят из науки как невостребованные временем и практикой.

Субъективная оценка позволяет также говорить о потере среди ученых интереса к строгому структурированию научных идей, систематизации знаний, аксиоматизации, к формированию на этой основе новых парадигм. Именно разный методологический подход к научному познанию отличает "творцов"-ученых от "ремесленников"-ученых. Такое отличие в непосредственном приложении к науке хорошо отражено в английских словах *researcher* и *scientist.*

Одновременно с потерей интереса к теоретическому подходу в исследованиях наблюдается снижение использования соответствующего ему строгого формализованного метода исследования. В целом можно говорить о потере интереса к дедуктивным методам познания. Для подтверждения этого приведем выдержку из статьи 2010 г. журнала "Nature Reviews. Microbiology". В этом издании, которое

определяет тенденции научных исследований в микробиологии, была опубликована статья под названием "Пирамида знаний" (*The pyramid of knowledge*) Д. Верникоса [5]. В статье он декларирует следующее: "дедуктивные методы обычно используются при недостатке рабочих данных и [они] не могут быть полностью аналитическими, тогда как индуктивный подход в конечном итоге приводит к более реалистичному пониманию общей картины" [5]. Эта выдержка демонстрирует позицию автора, который отдает предпочтение аналитическим методам, рассматривая метод синтеза и дедукции как нечто ошибочное и ущербное, способное лишь закрыть дыры научного знания в ожидании экспериментальных данных. В данной статье "правильное" научное познание описывается движением снизу вверх без особого плана и проекта, т. е. "на ощупь". Кроме того, в этой работе отражен текущий сдвиг науки от теории к эксперименту.

Изменение научной картины мира не обходится без последствий. Последствием нарушения ранее существовавшего равновесия между теорией и эмпирикой является фрагментация знаний, обрывочность и бессистемность научного поиска и менее глубокое погружение ученых в суть исследуемого объекта. Это и является причиной кризиса идеи, кризиса принятой ранее парадигмы.

Фундаментальные и прикладные науки в постканиталистическую эпоху. Вторая тенденция изменения науки, напрямую связанная с первой, заключается в том, что существующее смещение в исследованиях от фундаментальных к прикладным сопровождается их быстрым переходом на рельсы коммерциализации. Поскольку фундаментальная наука нуждается в долгосрочных инвестициях, она становится невыгодной в капиталистическом мире быстрых дивидендов. Такая инволюция в науке привела к изменению научной картины в целом. В 2015 г. в журнале "Nature" ученые формулируют проблему следующим образом: "Ученые, которые консультируют политиков, имеют два варианта: быть прагматичными или игнорируемыми" (*Scientists who advise politicians have two options: to be pragmatic or neglected*) [1].

Можно выделить следующие следствия перемещения основного акцента от фундаментальных исследований к прикладным и к их коммерческому применению. Во-первых, это потеря фундаментальной основы научных знаний по отношению к различным областям наук, которая объединяет и обеспечивает связь между этими областями. Во-вторых, происходит реконструкция базовых системообразующих парадигм познания и общих научных законов. Новые научные знания требуют постоянной эволюции и пересмотра их системообразующих парадигм, однако с учетом сдвига акцента исследований новые парадигмы не появляются. В-третьих, как следствие тенденции к коммерциализации науки и стремления к быстрой практической реализации результатов исследований, формируется потребительское отношение на всех уровнях научного сообщества, включая соответствующую такому отношению профессиональную этику. Данные изменения приводят к профессиональной стагнации и, как следствие, к утрате научного потенциала.

Кроме представленных выше тенденций, можно усмотреть признаки глобализации в науке. Помимо резко возросшей частоты использования слов с корнем "глобал", признаками глобализации в научном мире являются: утрата научных школ; появление головного научного центра с функцией цензора; формирование подчиненных транснациональных исследовательских центров, как правило, ограниченных экономически и юридически. Это обстоятельство приводит к эффектам застоя и самоторможения в науке, к определению ее статуса как вторичного, зависящего от практической ее востребованности.

Формальная постановка цели и задач исследования

Цель исследования, результаты которого представлены в настоящей работе: в первом приближении продемонстрировать тенденции изменения научного знания в отношении теории, практики, фундаментального и прикладного знания; продемонстрировать тенденции к коммерциализации науки и появлению элементов ее политизированности.

К задачам такого исследования относятся следующие:

1) предложить метод кластеризации документов с использованием заданного списка слов;

2) исходя из предположения, что научные тенденции коррелируют с изменением использования определенных слов и фраз в научных статьях, проанализировать опубликованные работы высокорейтингового в современной мировой науке журнала "Science" на период с 1996 по 2017 г.

Методы исследования

Для формирования корпуса текстов было проанализировано хранилище веб-ресурсов журнала "Science" (http://www.sciencemag.org) за 20 лет. При этом был реализован автоматизированный сбор статей скриптом, написанным на языке Python 3, с использованием библиотек Requests, urllib2 и selenium libraries с интеграцией браузеров Selenium и Firefox.

Предварительная стандартизация (токенизация, лемматизация, ликвидация стоп-слов) и формирование корпуса слов. В качестве предварительной подготовки к формированию корпуса слов была проведена

токенизация текстов (разбиение текста на более мелкие части — токены) на отдельные слова.

Далее был сформирован предварительный корпус из токенов (отдельных слов в различной падежной форме).

На следующем шаге корпус токенов был лемматизирован — изоформы (производные слов) были объединены в один лингвистический стержень (пул) путем сведения слов к определенной общей словарной корневой форме. Для этой цели использовали корпус WordNet® (http://wordnet.princeton.edu), английскую лексическую базу данных и библиотеку с открытым исходным кодом Natural Language Toolkit (NLTK) (http://www.nltk.org/) для программирования на языке Python 3. Корпус WordNet® состоит из 117 000 синсетов (когнитивных синонимов), объединенных в ограниченное число понятий (логически связанных групп терминов).

Из полученного корпуса лемматизированных слов убирали наиболее распространенные слова (стоп-слова) английского языка (местоимения, союзы и предлоги), которые были отфильтрованы с использованием библиотеки NLTK с корпусом из 128 английских стоп-слов.

На основе собранных текстов и полученного корпуса лемматизированных слов был сформирован используемый в анализе корпус слов размером 200 тыс. слов.

Формирование классов слов. На первом этапе из корпуса слов подбирали слова с нужным семантическим значением, т. е. слова, относящиеся к одной из следующих областей: фундаментальная наука; прикладная наука; теоретическое знание; практическое знание; коммерция; политика. Из этих слов создавались классы слов: "теоретический" и "фундаментальный" слово-классы; "прикладной" и "практический" слово-классы; "коммерческий" слово-класс; "политический" слово-класс.

Для создания перечисленных выше классов слов, во-первых, определялись центры этих классов, которые содержали по несколько ключевых слов, характеризующих каждую из областей: фундаментальная наука; прикладная наука; теоретическое знание; практическое знание; коммерция; политика. Во-вторых, основываясь на словах из центров слово-классов, в автоматическом режиме в корпусе полнотекстовых статей проводился поиск парных слов, которые появлялись со словами из центра слово-класса в одном предложении. Таким образом, каждое слово из центра слово-класса образовывало несколько сотен пар слов, в которых одно слово относилось к центру слово-класса, а другое — встречалось с этим словом в одном предложении во всем корпусе полнотекстовых статей. В-третьих, проводился учет совместной вероятности каждой такой пары слов с использованием для этого модели "Pointwise Mutual Information" [6]:

$$pm(c,\ w) = \ln\left(\frac{p(c,\ w)}{p(c)\,p(w)}\right),$$

где $p(c)$ — вероятность появления слова из центра слово-класса в полном корпусе статей; $p(w)$ — вероятность появления другого слова из пары в полном корпусе статей; $p(c,\ w)$ — вероятность нахождения обоих слов в одном предложении, рассчитанная на полном корпусе статей; $pm(c,\ w)$ — совместная вероятность слов центра $p(c)$ и слов первого круга $p(w)$. Здесь под словами первого круга для каждого слово-класса понимают слова, составляющие со словами из центра слово-класса пары с высокой совместной вероятностью.

В-четвертых, таким же образом, предварительно определив как центр слова первого круга, находили второй круг слов.

Разделить "теоретический" и "фундаментальный" классы не представлялось возможным, поэтому эти слово-классы были объединены в один слово-класс. По той же причине два слово-класса "прикладной" и "практический" были слиты в один.

После выполнения описанных выше расчетов получили кластеры слов, характеризующих классы: "фундаментальный" ("теоретический") слово-класс; "практический" ("прикладной") слово-класс; "политический" слово-класс и "коммерческий" слово-класс.

Векторная модель представления документа. Текстовый документ представлялся вектором слов, для формирования которого использовалась классическая *tf-idf*-модель (*term frequency inverse document frequency*) [3], позволяющая присвоить большой вес словам, которые встречались с высокой частотой в пределах одной статьи, но имели низкую частоту встречаемости в пределах всего корпуса статей. Таким образом, модель позволяла выделять слова, характерные именно для этой статьи, т. е. обладающие высокой специфичностью для отдельной статьи.

Мера *tf-idf* вычислялась по формуле

$$tfidf\,(t,\ d,\ D) = tf\,(t,\ d)\,idf\,(t,\ D),$$

где t — слово; d — статья в корпусе статей; D — корпус полнотекстовых статей.

В данной модели *tf* (*term frequency* — частота слова) — отношение числа появлений некоторого слова в статье к общему числу слов статьи, что показывало уровень специфики данного слова для статьи. Такая частота определялась по формуле

$$tf(t,d) = \frac{n_t}{\sum_k n_k},$$

где n_t — число появлений слова t в статье; в знаменателе — общее число слов в статье.

В используемой модели *idf* (*inverse document frequency* — обратная частота документа) — инверсия частоты, с которой некоторое слово встречается в корпусе статей. Эта величина уменьшает вес не обладающих спецификой, часто встречающихся в корпусе слов, и рассчитывается по формуле

$$idf(t,\ D) = \ln \frac{|D|}{\left|\{d_i \in D | t \in d_i\}\right|},$$

где |D| — число статей в корпусе статей; $\left|\{d_i \in D | t \in d_i\}\right|$ — число статей в корпусе статей, в которых встречается слово t.

Далее для каждой статьи в корпусе статей строили вектор значений *tf-idf* как по общему корпусу слов, так и отдельно по каждому слово-классу, так как задачей было не само разделение документов, а разделение документов согласно определенному принципу.

Метод главных компонент. Для понижения размерности пространства данных статей, представленных в виде вектора значений *tf-idf*, использовался анализ главных компонент. Полученные для статей векторы значений *tf-idf* размера 10...200 тыс. переводили в 5-мерное линейное многообразие методом главных компонент с использованием библиотеки Sklearn для Python 3 (http://scikit-learn.org/stable/index.html). Размерность была подобрана эмпирически с учетом последующей кластеризации, а именно: повышение и понижение размерности линейного многообразия не давало эффекта при последующей кластеризации методом *k*-средних — не улучшало разделение на два кластера.

Метод k-средних. Методом *k* -средних [4] кластеризовались данные статей, представленные по двум слово-классам: "фундаментальный" ("теоретический") слово-класс; "практический" ("прикладной") слово-класс. Отдельно данные кластеризовались по "политическому" слово-классу и "коммерческому" слово-классу. Кластеризация для каждой задачи проводилась не менее 5 раз. Была использована библиотека Scipy для Python 3, *k* выбирали равным 2 с возможностью пересечений кластеров.

Результаты

С использованием описанного выше автоматизированного режима сбора текстов был сформирован корпус полнотекстовых статей, опубликованных в журнале "Science" с 1996 по 2017 г. Корпус статей состоит из 16 тыс. научных публикаций (по 500...1500 слов каждая), хранящихся в виде отдельного текстового документа на локальном сервере.

Корпус слов формировался на основе лемматизированных слов. Объем корпуса составил 200 тыс. слов.

Предварительная визуализация данных. Простой анализ частоты слов показал, что частота употребления слов в научном тексте в публикациях журнала "Science" за последние 20 лет изменилась. Наиболее отчетливо это прослеживается на словах из "коммерческого" слово-класса.

На рис. 1 прослеживается постепенное увеличение частоты встречаемости слов из "коммерческого" слово-класса со временем, косвенно отражающее тенденцию к коммерциализации науки. Для примера были выделены следующие слова: стоимость (*cost*), прибыль (*benefit*), банк (*bank*), коммерция (*commerci*), инвестирование (*invest*), деньги (*money*), капитал (*capit*), фонд (*fund*), финансы (*financ*), грант (*grant*), прибыль (*profit*), доллар (*dollar*), евро (*euro*), продавец (*seller*), покупка (*buy*), покупатель (*buyer*). Слова приведены в лемматизированной форме.

Анализ главных компонент. С использованием анализа главных компонент были выделены основные компоненты для анализа статей, представленных в виде *tf-idf*-векторов.

На рис. 2 (см. вторую сторону обложки) можно наблюдать тенденцию к смещению употребления слов на примере полного корпуса слов "политического", "коммерческого" слово-классов, а также "фундаментального" ("теоретического") и "практического" ("прикладного") слово-классов.

Кластеризация документов на основе слово-классов. После кластеризации документов методом *k*-средних ($k = 2$) на основе векторов на базе "фундаментального" ("теоретического") и "практического" ("прикладного") слово-классов были получены два кластера достаточно хорошо (согласно принятой методике) разделяемых научных статей.

Первый кластер соответствовал теоретическим и фундаментальным статьям, его центр тяжести находился в области 1999—2000 гг. Второй кластер, соответствующий прикладным (практическим) научным статьям, имел центр тяжести в области 2010 г. (рис. 3, см. третью сторону обложки).

Кластеризация статей, представленных в виде векторов на основе "коммерческого" слово-класса, показала, что статьи могут быть разбиты на два пересекающихся кластера.

Кластер документов, характеризующийся превалированием слов из "коммерческого" и "политического" слово-классов, имел центр тяжести в области 2010 г., тогда как кластер статей "некоммерческой" и "деполитизированной" направленности имел центр тяжести в области 1999 г. (рис. 4, см. третью сторону обложки).

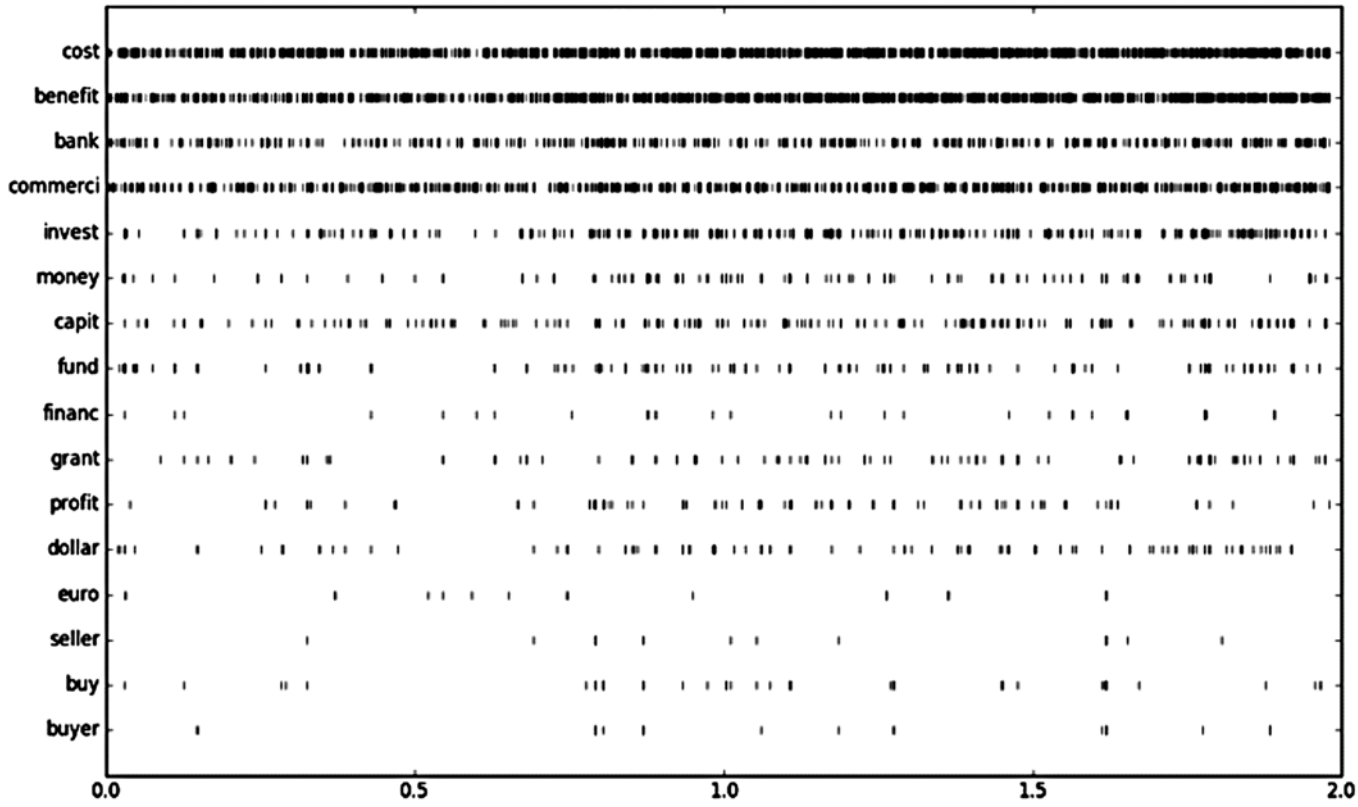

Рис. 1. Частотное распределение некоторых слов (приведенных к основной форме), входящих в "коммерческий" слово-класс по годам. На оси абсцисс 0 соответствует 1996 г., 2 — 2017 г. Каждой вертикальной черте соответствует появление слова в тексте. Текст был представлен последовательной склейкой статей в соответствии с их появлением с 1996 по 2017 г.

Заключение

В настоящей работе приведен новый метод кластеризации научных документов, позволяющий разделить текстовые документы, опираясь на определенные семантические потребности. В статье продемонстрированы данные, позволяющие заметить трансформацию научных публикаций с 1996 по 2017 г. на примере статей, опубликованных в журнале "Science". Если исходить из предположения, что журнал "Science" является одним из ведущих, высокорейтинговых журналов, отображающих состояние и тенденции мировой науки, и он вполне репрезентативен в контексте поставленной задачи, то рассуждения о состоянии, полученные на интуитивном уровне и изложенные в начале настоящей статьи, нашли некоторое фактическое подтверждение при использовании для анализа модели кластеризации документов на основе классов слов определенной направленности.

Таким образом, с высокой долей вероятности можно констатировать, что представленное во введении предположение о том, что научные публикации за последние 20 лет стали более политизированными и коммерциализированными, подтвердилось. Кроме того, показана тенденция смещения в последние 20 лет научного интереса в прикладные практические области при снижении числа публикаций, основанных на результатах исследований теоретического и фундаментального характера.

Список литературы

1. **Geden O.** Climate advisers must maintain integrity // Nature. 2015. Vol. 521 (7550). P. 27—28.

2. **Horgan J.** The End Of Science: Facing The Limits Of Knowledge In The Twilight Of The Scientific Age. Basic Books, 2015. 333 p.

3. **Jones K. S.** A statistical interepretation of tern specificity and its application in retrieval // Journal of Documentation. 2004. Vol. 60, N 5. P. 493—502.

4. **MacQueen J. B.** Some Methods for classification and Analysis of Multivariate Observations // In Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press. 1967. P. 281—297.

5. **Vernikos G. S.** The pyramid of knowledge // Nature Reviews. Microbiology. 2010. Vol. 8, N 2. P. 91.

6. **Zhai C. X., Massung S.** Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining. New York: Association for Computing Machinery and Morgan & Claypool, 2016. 531 p.

Scientific Text Analysis and New World Trends

A. Bernadotte, alexandra.bernadotte@gmail.com, Lomonosov Moscow State University, Moscow, 119991, Russian Federation; Department of Medical Biochemistry and Biophysics, Karolinska Institutet, S-171 77, Stockholm, Sweden

Corresponding author:

Bernadotte Alexandra, MD, PhD in Medicine, PhD candidate in Mathematics, Lomonosov Moscow State University, 119991, Moscow, Russian Federation; Adjunct Research Assistant Professor at Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, S-171 77, Sweden
E-mail: alexandra.bernadotte@gmail.com

Received on November 10, 2017
Accepted on November 28, 2017

Many of us of different fields and branches who are diving deep in science can notice slight or even quite sharp changes in the scientific world. The changes touch upon every part of science including semantics, methodology, axiomatic base, and objectives. From our experience one of the current scientific trends is a shift from theoretical science research to empirical research. The second trend is directly connected with the first one; there is a transformation from the basic science into an applied one, with a rapid transition to the commercial area. Being scientists, we also can notice marks of politicization and globalization in the scientific world.

To transform our vague senses and feelings into the scientifically-recognizable form, we analyzed scientific papers published during the period of 20 years in frames of the top scientific journal — Science. To reach theory-practice equilibrium shift and dominance of commercialization and politicization we analyzed text data, assuming that the loss of interest of theoretical and basic knowledge correlates with the usage of the certain words and phrases in scientific papers, as well as changes in their meaning and usages.

Indeed, recent work has demonstrated a 20-years transformation of scientific semantics and scientific interests. This paper has shown the growing commercialization and politicization of science, and reflected the current primacy of the application of knowledge upon basic science.

This work also introduces a new word-classes model of document clustering algorithms based on preliminarily words clustering with a usage of ontology-based term similarity, which can be helpful for text data mining and recommendation system.

Keywords: *text analysis, semantic analysis, tf-idf, scientific text, cognition, scientific trends, text clustering, data mining*

For citation:

Bernadotte A. Scientific Text Analysis and New World Trends, *Programmnaya Ingeneria,* 2018, vol. 9, no. 2, pp. 76—81.

DOI: 10.17587/prin.9.76-81

References

1. **Geden O.** Climate advisers must maintain integrity, *Nature*, 2015, vol. 521, no. 7550, pp. 27—28.

2. **Horgan J.** *The End of Science: Facing The Limits Of Knowledge In The Twilight Of The Scientific Age*, NY, Basic Books, 2015, 368 p.

3. **Jones K. S.** A statistical interepretation of tern specificity and its application in retrieval, *Journal of Documentation*, 2004, vol. 60, no. 5, pp. 493—502.

4. **MacQueen J. B.** Some Methods for classification and Analysis of Multivariate Observations, *5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 1967, pp. 281—297.

5. **Vernikos G. S.** The pyramid of knowledge, *Nature Reviews. Microbiology*, 2010, vol. 8, no. 2, pp. 91.

6. **Zhai C. X., Massung S.** *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining,* NY, Association for Computing Machinery and Morgan & Claypool, 2016, 531 p.

УДК 004.855.5 DOI: 10.17587/prin.9.82-90

В. А. Харахинов, аспирант, e-mail: tes4obse@mail.ru,
С. С. Сосинская, канд. техн. наук, доц., проф., e-mail: sosinskaya@mail.ru,
Иркутский национальный исследовательский технический университет

Влияние сокращения размерности пространства признаков на результаты классификации листьев различных видов растений

Рассмотрен процесс проведения классификации листьев различных видов растений, описываемых набором числовых признаков, на основе использования многослойного персептрона.

Предложено применение методов факторного анализа в целях сокращения размерности исходного пространства признаков. Для определения числа факторов применены два популярных критерия: критерий Кайзера и критерий доли воспроизводимой дисперсии.

В целях уменьшения временных затрат на обучение сетей выполнена классификация выборки листьев как в исходном пространстве признаков, так и в пространстве координат факторов.

Качество классификации и временные затраты в процессе обучения сети в том и другом случае отображено в виде таблиц. Графики, отражающие уменьшение временных затрат на обучение сетей при использовании факторных координат и зависимость процента ошибок классификации от числа факторов, позволяют сделать вывод о том, какая сеть при условии правильного подбора числа факторов дает достаточно эффективный способ уменьшения временных затрат в процессе обучения сети при достижении достаточно высокого качества классификации.

Ключевые слова: *классификация, сети прямого распространения, машинное обучение, факторный анализ*

Введение

К настоящему времени разработано большое число методов для выполнения классификации данных, к которым относятся: классификация на основе деревьев решений, байесовская классификация, классификация методом опорных векторов, а также ряд других. В последние годы для решения задач классификации активно применяют нейросетевой подход, основное преимущество которого — возможность настройки параметров в зависимости от объектов обучающей выборки.

Несмотря на отмеченное преимущество нейронных сетей, они имеют ряд недостатков, например, от обучающей выборки зависит архитектура сети, а именно — число слоев, число нейронов и функция активации каждого слоя. Процесс обучения сети может выполняться достаточно медленно вследствие большой размерности входных данных.

В настоящей статье рассмотрены возможности многослойных персептронов и методы факторного анализа для сокращения размерности данных при проведении классификации листьев различных пород растений. Приведены затраты времени на обучение сети и качество распознавания на контролирующих данных, описывающих заданную выборку, как в исходном признаковом пространстве, так и в пространстве координат факторов.

Классификация

Под классификацией объектов (наблюдений, событий), описываемых набором числовых признаков, понимают способы отнесения этих объектов к одному из заранее известных классов. Классификация относится к стратегии обучения с учителем, которое также именуют контролируемым или управляемым обучением. С математической точки зрения процесс классификации можно описать следующим образом:

Пусть X — множество описаний объектов; Y — конечное множество номеров (имен, меток) классов. Существует неизвестная целевая зависимость — ото-

бражение $y^*: X \to Y$, значения которой известны только на объектах конечной обучающей выборки $X^m = \{(x_1, y_1), \ldots, (x_m, y_m)\}$. Требуется построить алгоритм $\alpha: X \to Y$, способный классифицировать произвольный объект $x \in X$ [1].

В работе для проведения классификации листьев различных пород растений использован нейросетевой подход, в основе которого — многослойный персептрон.

Исходные данные

Исходная выборка содержит 340 объектов, разделенных на 30 классов. Каждый объект описывается 14 признаками. Признаки 1—8 описывают форму листка растения, включающую: эксцентричность; соотношение сторон листка; элонгацию; цельность; стохастическую выпуклость; изометрический фактор; максимальную глубину отпечатка; дольчатость. Признаки 9—14 описывают цифровое изображение листка и включают: среднюю интенсивность; среднюю контрастность; гладкость; момент третьего порядка; однородность; энтропию. Данные получены из работы [2]. Эта выборка интересна для проводимого анализа в связи с большим числом классов и сравнительно большим числом признаков. Однако ее недостатком является не очень большой объем: каждый класс содержит в среднем по 11 объектов, поэтому отбор для обучения части выборки неизбежно ухудшит результат обучения, в связи с чем разделение исходной выборки на обучающую и тестирующую не выполнялось, т. е. обучение и последующую классификацию проводили на полной выборке.

Использование нейронных сетей для классификации

Реализация нейросетевого подхода к решению задачи классификации листьев различных растений была выполнена в среде системы MATLAB с использованием пакета расширения Neural Network Toolbox [3].

В многослойной нейронной сети нейроны располагаются по слоям. Нейроны первого слоя получают входные сигналы, преобразуют их и передают нейронам второго слоя, далее срабатывает второй слой, и т. д. до последнего слоя, который выдает выходные сигналы. Такая сеть называется многослойной сетью прямого распространения, или многослойным персептроном.

Промежуточные слои между внешним входным сигналом и выходным слоем называют скрытыми.

На рис. 1 показан граф многослойного персептрона с двумя скрытыми слоями.

Обучение многослойного персептрона осуществляется с помощью алгоритма обратного распространения ошибки и его модификаций. В данной работе для обучения сети использовался квазиньютоновский алгоритм, а именно алгоритм Левенберга—Марквардта, который предназначен для оптимизации параметров нелинейных регрессионных моделей [4]. Предполагается, что в качестве критерия оптимизации используется средняя квадратичная ошибка модели на обучающей выборке. Алгоритм заключается в последовательном приближении заданных начальных значений параметров к искомому локальному оптимуму [5].

С математической точки зрения алгоритм Левенберга—Марквардта описывается следующим образом.

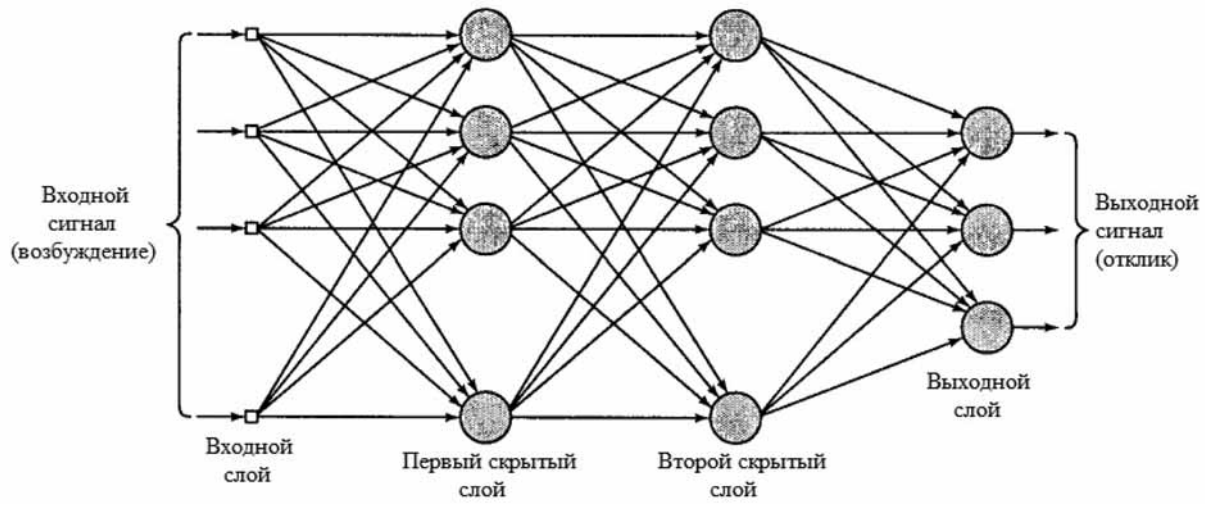

Рис. 1. Граф многослойного персептрона с двумя скрытыми слоями

Задана обучающая выборка — множество пар $D=\{(x_n, y_n)\}_{n=1}^{N}$ свободной переменной $x \in R^M$ (входы сети) и зависимой переменной $y \in R$. Задана функциональная зависимость, представляющая собой регрессионную модель $y=f(w, x_n)$, непрерывно дифференцируемая в области $W \times X$. Требуется найти такое значение вектора параметров w, которое бы доставляло локальный минимум функции ошибки

$$E_D=\sum_{n=1}^{N}\left(y_n-f(w, x_n)\right)^2.$$

Перед началом работы алгоритма задается начальный вектор параметров w. На каждом шаге итерации этот вектор заменяется на вектор $w+\Delta w$. Для оценки приращения Δw используется линейное приближение функции $f(w+\Delta w, x)\approx f(w, x)+\mathbf{J}\Delta w$, где $\mathbf{J}$ — якобиан функции $f(w, x_n)$. Матрицу $\mathbf{J}$ размером $N \times R$ можно представить в виде

$$\mathbf{J}=\begin{bmatrix}\frac{\partial f(w, x_1)}{\partial w_1} & \cdots & \frac{\partial f(w, x_1)}{\partial w_R}\\ \vdots & \ddots & \vdots\\ \frac{\partial f(w, x_N)}{\partial w_1} & \cdots & \frac{\partial f(w, x_N)}{\partial w_R}\end{bmatrix}.$$

Здесь вектор параметров $w=[w_1,\cdots,w_R]^{\mathbf{T}}$.

Чтобы найти значение Δw нужно решить систему линейных уравнений $\Delta w=\left(\mathbf{J}^{\mathbf{T}}\mathbf{J}\right)^{-1}\mathbf{J}^{\mathbf{T}}\left(y-f(w)\right)$. Так как число обусловленности матрицы $\mathbf{J}^{\mathbf{T}}\mathbf{J}$ есть квадрат числа обусловленности матрицы $\mathbf{J}$, то матрица $\mathbf{J}^{\mathbf{T}}\mathbf{J}$ может оказаться существенно вырожденной. Поэтому Марквардт предложил ввести параметр регуляризации $\lambda \geqslant 0$, следовательно, $\Delta w=\left(\mathbf{J}^{\mathbf{T}}\mathbf{J}+\lambda\mathbf{I}\right)^{-1}\mathbf{J}^{\mathbf{T}}\left(y-f(w)\right)$.

В данном случае $\mathbf{I}$ — единичная матрица. Параметр λ назначается на каждой итерации алгоритма. Если значение ошибки E_D убывает быстро, малое значение λ сводит этот алгоритм к алгоритму Гаусса—Ньютона [6]. Выбор этого алгоритма для обучения нейронной сети обусловлен тем, что метод сходится за меньшее число итераций в сравнении с другими квазиньютоновскими алгоритмами [7].

Для экспериментов использовали функцию формирования многослойного персептрона с различным числом слоев, реализованную одним из авторов, так как такой функции нет в MATLAB. На рис. 2 и 3 приведена архитектура спроектированной многослойной сети для классификации.

Как видно на рис. 2, сеть имеет один скрытый слой, в котором используется сигмоидальная функция активации. Число нейронов в этом слое должно быть не меньше числа нейронов в выходном слое. Выходной слой использует линейную функцию активации. Число нейронов во входном слое равно числу признаков, описывающих объект классификации. Число нейронов в выходном слое равно числу классов в исходной выборке.

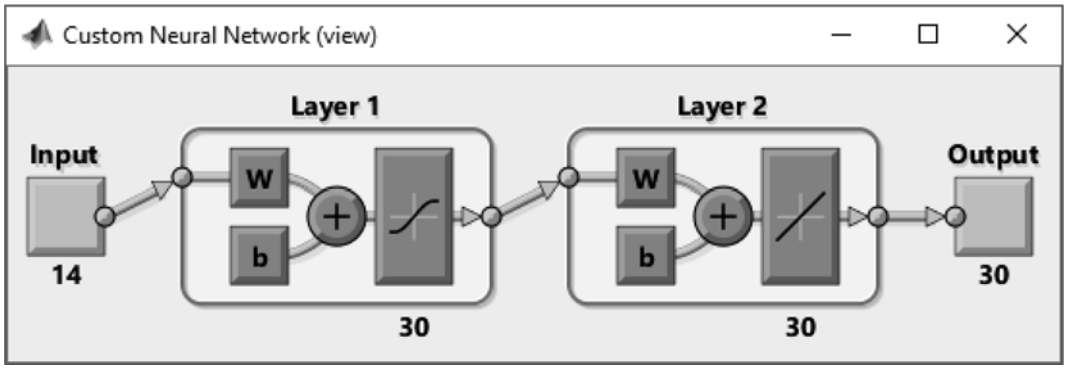

Рис. 2. Архитектура двухслойной сети в среде MATLAB

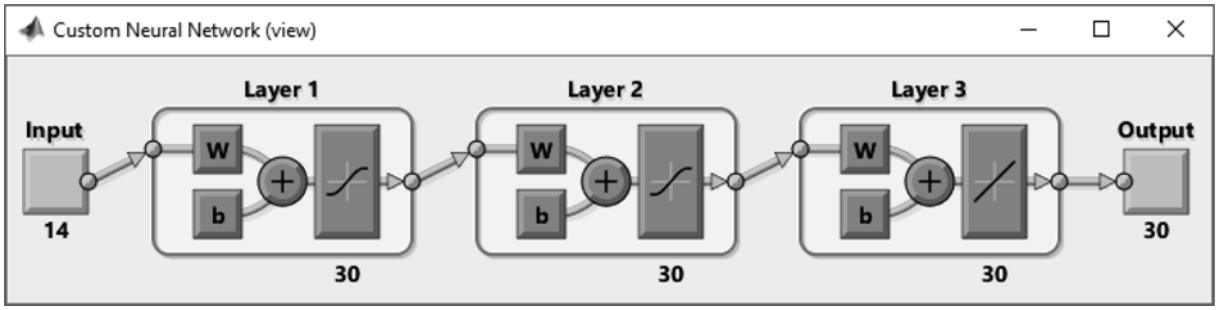

Рис. 3. Архитектура трехслойной сети в среде MATLAB

На рис. 3 показана трехслойная сеть, которая имеет два скрытых слоя, использующих сигмоидальную функцию активации. В остальном она схожа с двухслойной сетью.

Результаты классификации

Табл. 1 содержит данные о времени и качестве классификации листьев различных видов растений.

Для оценки качества работы классификатора использовалась функция Matlab confusionmat, которая подсчитывает матрицу различий между известными и предсказанными номерами классов для классифицируемых точек. Эти значения затем пересчитывались в проценты и усреднялись по всем классам.

На основе данных табл. 1 можно сделать вывод, что для используемой выборки трехслойный персептрон дает лучшее качество классификации, чем двухслойный.

Многослойный персептрон обладает высокой степенью связности, благодаря чему уменьшение числа нейронов во входном слое приведет к значительному уменьшению затрат времени на обучение сети. Для этого используются методы понижения размерности данных, в частности, факторный анализ.

Факторный анализ

В факторном анализе предполагается, что значения исходных измеряемых переменных приближенно находятся в линейной зависимости от факторов. Другими словами — если бы мы знали факторы, то могли бы рассчитать исходные переменные по формуле $\mathbf{X} = \mathbf{AF} + \mathbf{U}$, где $\mathbf{X}$ —нормализованный входной вектор; $\mathbf{A}$ — матрица определяемых факторных нагрузок; $\mathbf{F}$ — вектор общих факторов; $\mathbf{U}$ —вектор характерных факторов. Общие факторы называют так потому, что они едины для всех переменных. В отличие от них характерный фактор для каждой переменной свой, так как он определяется природой данной переменной и по сути имеет смысл "помехи".

Таблица 1

Результаты классификации по исходным данным (алгоритм обучения Левенберга—Марквардта)

Тип сети	Число итераций на обучение	Время на обучение, с	Процент ошибок
Персептрон (два слоя)	40	33,5225	8,5294
	80	66,5639	3,5294
	160	131,9929	2,0588
	240	200,0913	3,5294
Персептрон (три слоя)	20	55,4973	2,9412
	40	108,5084	0

Размер матрицы $\mathbf{A}$ равен $d \times m$, где d — число признаков объектов; m — число факторов. Элемент матрицы $A_{i,j}$ называется нагрузкой i-й переменной на j-й фактор, или наоборот, нагрузкой j-го фактора на i-ю переменную. Число элементов векторов $\mathbf{X}$, $\mathbf{F}$ и $\mathbf{U}$ равно d [8].

Каждый фактор зависит от сильно коррелирующих между собой признаков. Как следствие, происходит перераспределение дисперсии между признаками, в результате чего получается максимально простая и наглядная структура факторов. В этих целях была рассчитана корреляционная матрица для исходной выборки (табл. 2).

Важной составляющей факторного анализа является процедура вращения факторов, т. е. перераспределения дисперсии по определенному методу.

Нагрузки, полученные в качестве одного из результатов факторного анализа, когда число факторов больше единицы, неоднозначны, и существует возможность получать различные эквивалентные множества нагрузок путем вращения факторной структуры.

Цель ортогональных вращений — определение простой структуры факторных нагрузок [9]. Решение общей задачи вращения (для ортогональных факторов) можно записать в следующем виде: $\mathbf{B} = \mathbf{AT}$, где $\mathbf{A} = (a_{jp})$ — исходная факторная матрица; $\mathbf{B} = (b_{jp})$ — финальная факторная матрица; $\mathbf{T} = (tb_{qp})$ — ортогональная матрица преобразования.

Наиболее часто используемым вычислительным методом вращения является varimax, предложенный Кайзером [10], который максимизирует разброс квадратов нагрузок для каждого фактора, что приводит к увеличению больших и уменьшению малых значений факторных нагрузок. Согласно Кайзеру простота фактора p определяется дисперсией квадратов его нагрузок, т. е.:

$$s_p^2 = \frac{1}{n}\sum_{j=1}^{n}\left(b_{jp}^2\right)^2 - \frac{1}{n^2}\left(\sum_{j=1}^{n} b_{jp}^2\right)^2 \quad (p = 1, 2, \ldots, m).$$

Если эта дисперсия максимальна, то фактор наилучшим образом интерпретируем, поскольку при этом его нагрузки близки в основном к единице или к нулю [9]. Критерий максимизации простоты полной факторной матрицы сводится, следовательно, к максимизации суммы значений по всем факторам:

$$s^2 = \sum_{p=1}^{m} s_p^2.$$

Критерий varimax можно описать в следующем виде [9]:

$$V = n\sum_{p=1}^{m}\sum_{j=1}^{n}\left(\frac{b_{jp}}{n_j}\right)^4 - \sum_{p=1}^{m}\left(\sum_{j=1}^{n}\frac{b_{jp}^2}{n_j^2}\right)^2.$$

Таблица 2

Корреляционная матрица по исходным данным

Признаки	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1,0000	0,5511	0,5536	0,3735	0,3863	−0,0361	−0,2744	−0,2136	−0,2263	−0,1947	−0,1911	−0,1496	−0,2503	−0,2397
2	0,5511	1,0000	0,6783	0,0052	0,1071	−0,4710	0,0919	0,1224	−0,2822	−0,2975	−0,2633	−0,2338	−0,2294	−0,3129
3	0,5536	0,6783	1,0000	−0,4114	−0,3788	−0,7931	0,4371	0,4079	−0,2045	−0,1885	−0,1785	−0,1511	−0,2507	−0,2109
4	0,3735	0,0052	−0,4114	1,0000	0,8625	0,7555	−0,8857	−0,8238	0,0849	0,0827	0,0776	0,0589	0,1263	0,0556
5	0,3863	0,1071	−0,3788	0,8625	1,0000	0,6559	−0,7691	−0,6996	0,0632	0,0580	0,0507	0,0366	0,1067	0,0526
6	−0,0361	−0,4710	−0,7931	0,7555	0,6559	1,0000	−0,7358	−0,6242	0,0806	0,0782	0,0746	0,0665	0,1331	0,0698
7	−0,2744	0,0919	0,4371	−0,8857	−0,7691	−0,7358	1,0000	0,9465	−0,0905	−0,0887	−0,0590	−0,0232	−0,1378	−0,1117
8	−0,2136	0,1224	0,4079	−0,8238	−0,6996	−0,6242	0,9465	1,0000	−0,1691	−0,1778	−0,1490	−0,1192	−0,1692	−0,1811
9	−0,2263	−0,2822	−0,2045	0,0849	0,0632	0,0806	−0,0905	−0,1691	1,0000	0,9548	0,9554	0,8217	0,8045	0,9357
10	−0,1947	−0,2975	−0,1885	0,0827	0,0580	0,0782	−0,0887	−0,1778	0,9548	1,0000	0,9796	0,9179	0,6527	0,8601
11	−0,1911	−0,2633	−0,1785	0,0776	0,0507	0,0746	−0,0590	−0,1490	0,9554	0,9796	1,0000	0,9510	0,6392	0,8204
12	−0,1496	−0,2338	−0,1511	0,0589	0,0366	0,0665	−0,0232	−0,1192	0,8217	0,9179	0,9510	1,0000	0,4093	0,6450
13	−0,2503	−0,2294	−0,2507	0,1263	0,1067	0,1331	−0,1378	−0,1692	0,8045	0,6527	0,6392	0,4093	1,0000	0,7978
14	−0,2397	−0,3129	−0,2109	0,0556	0,0526	0,0698	−0,1117	−0,1811	0,9357	0,8601	0,8204	0,6450	0,7978	1,0000

Для проведения факторного анализа в системе MATLAB используется функция factoran, входящая в пакет расширения Statistics.

При вызове этой функции необходимо определить число факторов. Максимально возможное число простых факторов определяется неравенством $(d+m) \leqslant (d-m)^2$, где d — число признаков, m — число факторов [11]. В данном случае $d = 14$, следовательно, $1 \leqslant m \leqslant 9$.

Основной сложностью факторного анализа является выделение и интерпретация главных факторов. Существует несколько часто употребляемых критериев определения числа факторов. Некоторые из них являются альтернативными по отношению к другим, а часть этих критериев можно использовать совместно с тем, чтобы один дополнял другой.

В контексте данной статьи для определения числа факторов используются следующие два критерия.

1. Критерий Кайзера или критерий собственных чисел. Этот критерий предложен Кайзером, и является, вероятно, наиболее широко используемым. Отбираются только факторы с собственными значениями, равными или большими 1. Это означает, что если фактор не выделяет дисперсию, эквивалентную, по крайней мере, дисперсии одной переменной, то он опускается [12]. Собственные значения факторов вычисляются на основе значений корреляционной матрицы (см. табл. 2). В табл. 3 приведены собственные значения для каждого фактора.

Исходя из полученных значений, оптимальное число факторов, по критерию Кайзера, равно 3 ($m = 3$).

2. Критерий доли воспроизводимой дисперсии. Факторы ранжируются по доле воспроизводимой дисперсии, когда процент дисперсии оказывается несущественным выделение следует остановить [12]. В табл. 4 приведены проценты воспроизводимой дисперсии для каждого фактора.

Исходя из полученных значений, оптимальное число факторов, по критерию доли воспроизводимой дисперсии, равно 5 ($m = 5$). Процент общей дисперсии при $m = 5$ равен 99,9169 %.

Таблица 3

Собственные значения факторов

Фактор	Собственное значение
1	5,6829
2	4,1948
3	2,1021
4	0,7355
5	0,4377
6	0,3888
7	0,1712
8	0,1138
9	0,0734
10	0,0453
11	0,0246
12	0,0175
13	0,0121
14	0,0002

Таблица 4

Проценты воспроизводимой дисперсии факторов

Фактор	Дисперсия, %
1	82,1774
2	13,3138
3	3,6406
4	0,4372
5	0,3478
6	0,0385
7	0,0203
8	0,0131
9	0,0099
10	0,0008
11	0,0005
12	0,0000
13	0,0000
14	0,0000

Эксперименты с данными в координатах факторов

Использование данных в координатах факторов позволяет уменьшить число нейронов входного слоя сети. На рис. 4 приведена архитектура двухслойного персептрона при $m = 3$.

Табл. 5 содержит данные о времени и качестве классификации для данных, описанных в координатах факторов.

Анализируя эти результаты можно сделать вывод, что при сокращении исходного пространства признаков с $d = 14$ до $m = 3$ использование матрицы, содержащей полученные факторные значения в качестве обучающей выборки, является нерелевантным, поскольку заметно увеличивается процент ошибочных классификаций для любой из использованных сетей.

При сокращении исходного пространства признаков до $m = 5$ использование матрицы факторных

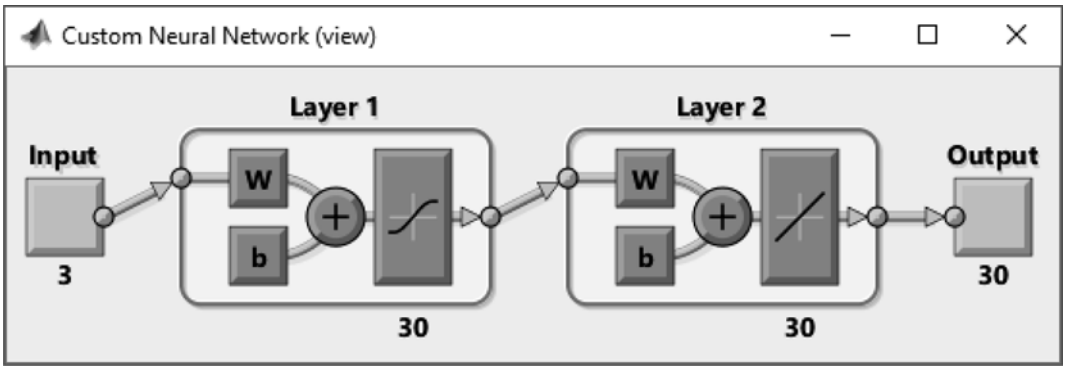

Рис. 4. Архитектура двухслойного персептрона при использовании факторных значений

Таблица 5

Результаты классификации по факторным значениям (алгоритм обучения Левенберга–Марквардта)

Тип сети	Число факторов	Число итераций на обучение	Время на обучение, с	Процент ошибок
Персептрон (два слоя)	3	40	15,5262	33,2353
		80	30,7103	30,8824
		160	61,2162	30
		240	90,9560	35,8824
	5	40	18,4330	23,5294
		80	36,2435	21,7647
		160	72,8267	20,5882
		240	108,1011	20,5882
Персептрон (три слоя)	3	20	39,9323	22,0588
		40	77,9125	11,7647
		60	115,6585	10,2941
	5	20	43,9026	4,4118
		40	85,7456	1,4706
		60	127,8268	2,3529

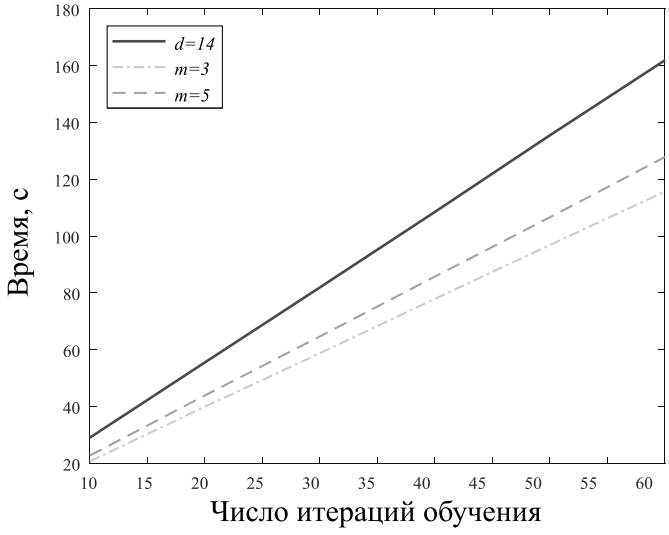

Рис. 5. График временных затрат на обучение трехслойной сети

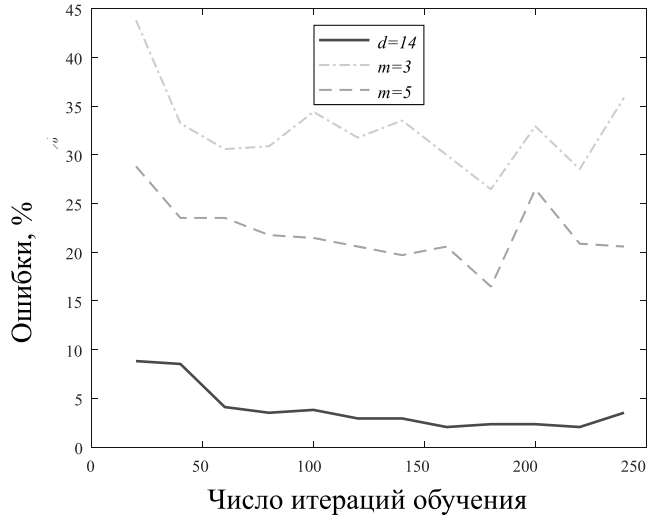

Рис. 6. Графическое представление результатов классификации при использовании двухслойной сети

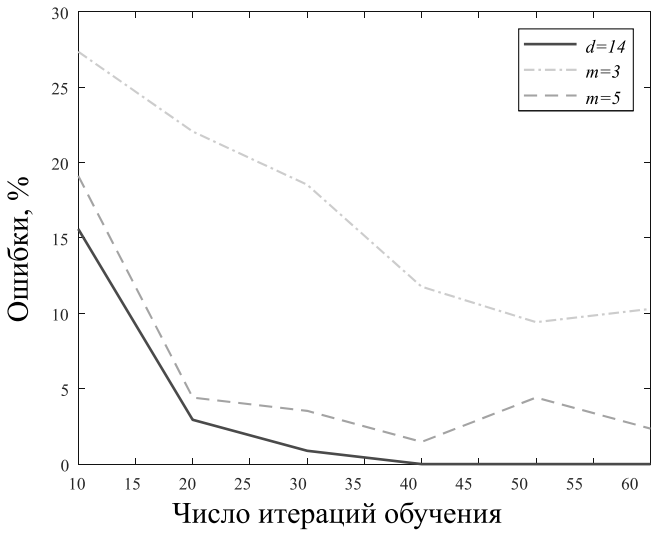

Рис. 7. Графическое представление результатов классификации при использовании трехслойной сети

значений в процессе классификации положительно сказывается на качестве распознавания трехслойного персептрона, сокращая время обучения и одновременно сохраняя сравнительно низкий процент ошибочных результатов.

На рис. 5 показано уменьшение временных затрат на обучение сети при использовании факторных значений.

На рис. 6 показана зависимость процента ошибочно классифицированных наблюдений от числа итераций при обучении двухслойной сети на различных вариантах данных: исходные данные (d = 14); факторные значения (m = 3 и m = 5).

На рис. 7 (по аналогии с рис. 6) показана зависимость процента ошибок при классификации от числа итераций при обучении трехслойной сети на различных вариантах данных: исходные данные (d = 14); факторные значения (m = 3 и m = 5).

Заключение

Анализ результатов исследований, представленных в данной статье, показывает, что нейронные сети прямого распространения могут быть использованы в процессах классификации данных в коллекциях, по своим характеристикам близких к тем, которые использовались для классификации листьев различных видов растений. Такой подход позволяет добиться низкого процента ошибок при классификации, однако следует отметить, что он влечет за собой значительные временные затраты.

Использование факторного анализа для сокращения исходного пространства признаков и выбор одного из двух критериев для определения числа факторов (критерия Кайзера и критерия доли воспроизводимой дисперсии) позволили выбрать наилучшее число факторов при применении второго критерия.

Таким образом, применение трехслойной сети, использующей матрицу данных в координатах факторов при условии правильного подбора их числа, — достаточно эффективный способ уменьшения временных затрат в процессе обучения сети, позволяющий добиться высокого качества классификации.

Представленная в данной статье методика исследований, на взгляд авторов, может применяться и для данных из других предметных областей, хотя результаты, вероятнее всего, в значительной степени будут зависеть от объема и характера выборок, числа признаков и количества классов.

Список литературы

1. **Айвазян С. А., Бухтштабер В. М., Енюков И. С., Мешалкин Л. Д.** Прикладная статистика: классификация и снижение размерности. М.: Финансы и статистика, 1989. 607 с.

2. **Репозиторий** реальных и модельных задач машинного обучения. URL: http://archive.ics.uci.edu/ml/datasets/Leaf

3. **Дьяконов В. П., Круглов В. В.** MATLAB 6.5 SP1/7/7 SP1/7 SP2 + Simulink 5/6. Инструменты искусственного интеллекта и биоинформатики. М.: СОЛОН-ПРЕСС, 2009. URL: http://www.studmedlib.ru/book/5-98003-255-X.html

4. **Хайкин С.** Нейронные сети: полный курс, 2-е издание. М.: Вильямс, 2006. 1104 с.

5. **Портал** искусственного интеллекта. URL: http://neuronus.com/theory/246-algoritm-levenberga-markvardta.html

6. **Демиденко Е. З.** Оптимизация и регрессия. М.: Наука, 1989. 296 с.

7. **Медведев В. С., Потемкин В. Г.** Нейронные сети. MATLAB 6. М.: ДИАЛОГ-МИФИ, 2002. 496 с.

8. **MATLAB** Documentation. URL: http://mathworks.com/help/stats/factoran.html

9. **Харман Г.** Современный факторный анализ. М.: Статистика, 1972. 484 с.

10. **Kaiser H. F.** The varimax criterion for analytic rotation in factor analysis // Phychometrika. 1958. Vol. 23. P. 187—200.

11. **Лоули Д., Мансвелл А.** Факторный анализ как статистический метод. М.: Мир, 1967. 141 с.

12. **Ким О. Дж., Мьюллер Ч. У., Клекка У. Р.** Факторный, дискриминантный и кластерный анализ. М.: Финансы и статистика, 1989. 215 с.

The Effect of Dimension Reducing on Classification Results of Leaves of Various Plant Species

V. A. Kharakhinov, e-mail: tes4obse@mail.ru, **S. S. Sosinskaya,** e-mail: sosinskaya@mail.ru, National Research Irkutsk State Technical University, Irkutsk, 664074, Russian Federation,

Corresponding author:

Sosinskaya Sophia S., Professor, National Research Irkutsk State Technical University, Irkutsk, 664074, Russian FederationE-mail: sosinskaya@mail.ru

Received on November 25, 2017
Accepted on December 04, 2017

The article considers the process of classification of different plant species, described by a set of numeric properties, based on multilayer perceptron.

It was proposed to apply the factor analysis to reducing dimension of the original data set. Two well-known criteria for determining the number of factors was used. Such as Kaiser criteria and variance explained criteria.

The neural networks were trained on native data set and on factor scores to compare time expenditures of training process.

The quality of classifications for both cases is displayed in tables. The graph displays time expenditure decreasing for network training when factor scores were used. Also the graphs display relation of classification error rates from a number of factors. These graphs allow one to conclude which network, with proper selection of the number of factors, provide an efficient way to reduce the time expenditure of training process, at the same time allow achieving a high quality of classification.

Keywords: *classification, feedforward network, machine learning, factor analysis*

For citation:

Kharakhinov V. A., Sosinskaya S. S. The Effect of Dimension Reducing on Classification Results of Leaves of Various Plunt Species, *Programmnaya Ingeneria,* 2018, vol. 9, no. 2, pp. 82—90.

DOI: 10.17587/prin.9.82-90

References

1. **Ajvazjan S. A., Buhtshtaber V. M., Enjukov I. S., Meshalkin L. D.** *Prikladnaja statistika: klassifikacija i snizhenie razmernosti* (Applied statistics: classification and reduction of dimensionality), Moscow, Finansy i statistika, 1989, 607 p. (in Russian).

2. **Machine** Learning Repository, available at: http://archive.ics.uci.edu/ml/datasets/Leaf

3. **Dyakonov V. P., Kruglov V. V.** *MATLAB 6.5 SP1/7/7 SP1/7 SP2 + Simulink 5/6. Instrumenty iskusstvennogo intellekta i bioinformatiki* (MATLAB 6.5 SP1/7/7 SP1/7 SP2 + Simulink 5/6. Instruments of artificial intelligence and bioinformatics), Moscow, SOLON-PRESS, 2009, available at: http://www.studmedlib.ru/book/5-98003-255-X.html

4. **Hajkin S.** *Nejronnye seti: polnyj kurs* (Neural networks: full course), 2nd ed., Moscow, Williams, 2006, 1104 p. (in Russian).

5. **Portal** iskusstvennogo intellekta (The portal of artificial intelligence), available at: http://neuronus.com/theory/246-algoritm-levenberga-markvardta.html (in Russian).

6. **Demidenko E. Z.** *Optimizacija i regressija* (Optimization and regression), Moscow, Nauka, 1989, 296 p. (in Russian).

7. **Medvedev V. S., Potemkin V. G.** *Nejronnye seti. MATLAB 6* (Neural network. MATLAB 6), Moscow, DIALOG-MIFI, 2002, 496 p. (in Russian).

8. **MATLAB** Documentation, available at: http://mathworks.com/help/stats/factoran.html

9. **Harman G.** *Sovremennyj faktornyj analiz* (Modern factor analysis.), Moscow, Statistika, 1972, 484 p. (in Russian).

10. **Kaiser H. F.** The varimax criterion for analytic rotation in factor analysis, *Phychometrika*, 1958, vol. 23, pp. 187—200.

11. **Louli D., Mansvell A.** *Faktornyj analiz kak statisticheskij metod* (Factor analysis as a statistical method), Moscow, Mir, 1967, 141 p. (in Russian).

12. **Kim O. D., M'juller Ch. U., Klekka U. R.** *Faktornyj, diskriminantnyj i klasternyj analiz* (Factorial, discriminant and cluster analysis), Moscow, Finansy i statistika, 1989, 215 p. (in Russian).

УДК 004.4 DOI: 10.17587/prin.9.91-96

С. А. Леоновец, ст. инженер[1], аспирант[2], e-mail: ser2694@ya.ru,
А. В. Гурьянов[1], канд. экон. наук, ген. дир., e-mail: postmaster@elavt.spb.ru,
А. В. Шукалов, канд. техн. наук, доц., первый заместитель генерального директора — главный конструктор[1], доц.[2], e-mail: aviation78@mail.ru,
И. О. Жаринов, д-р техн. наук, проф., руководитель учебно-научного центра[1], зав. кафедрой[2], e-mail: igor_rabota@pisem.net
[1]АО "ОКБ "Электроавтоматика", г. Санкт-Петербург,
[2]Университет ИТМО, г. Санкт-Петербург

Программное средство для автоматизации контроля жизненного цикла текстовой документации на программно-управляемые изделия

Рассмотрена задача автоматизации процесса подготовки, хранения и мониторинга контроля версий текстовой конструкторской и программной документации с помощью специализированного программного обеспечения. Автоматизация подготовки документации основана на обработке инженерных данных, которые содержатся в спецификациях и технической документации. Обработка данных предполагает наличие строго структурированных электронных документов, подготовленных в распространенных форматах в соответствии с шаблонами на основе отраслевых стандартов, и генерацию с помощью автоматизированного метода текстового технического документа. Контролируется жизненный цикл документа и технические данные, содержащиеся в нем. На каждом этапе жизненного цикла выполняется архивное хранение данных. Представлены результаты исследования по оценке быстродействия текстового редактора при использовании различных широко распространенных форматов документов для этапов их автоматизированного контроля и хранения. Описано новое разработанное программное обеспечение и инструментальные средства на его основе, облегчающие документирование результатов разработки бортового приборного оборудования.

Ключевые слова: *конструкторская документация, автоматизация, текстовые документы, жизненный цикл документа*

Введение

Результаты разработки бортового приборного оборудования в области авиационного приборостроения представляются в виде конструкторских и программных документов в соответствии с ГОСТ 2.102—68 и ГОСТ 19.101—77 на всех этапах проектирования: эскизное проектирование, технический проект, технические предложения, рабочее конструкторское проектирование. Эффективность деятельности организации, занимающейся проектными разработками, во многом зависит от времени подготовки и качества документации. Чтобы минимизировать время разработки документов, а также повысить качество документации, необходимо снизить число ошибок в документации, обусловленных, например, влиянием человеческого фактора. Это достигается автоматизированным управлением жизненным циклом документации и содержащихся в ней инженерных данных. На настоящее время существуют различные программные средства, которые обеспечивают автоматизированную поддержку разработки и контроля версий конструкторских и программных документов [1—4]. Однако они не поддерживают нормы на оформление и на структуру представления технической документации, изложенные в действующих государственных стандартах, например, в ГОСТ 2.105—95 и ГОСТ 2.104—68 [5, 6].

Исходя из изложенного выше, разработка отраслевых систем автоматизированного проектирования, которые позволяют избежать "ручного" контроля данных и осуществлять автоматизированную поддержку процесса сквозного проектирования конструкторской и программной документации на изделие, является актуальной. Основные концепции проектирования отраслевой САПР в области авиационного приборостроения подробно рассмотрены в работах [7—14].

Цель настоящей статьи заключается в представлении широкому кругу читателей результатов выполне-

ния составной части опытно-конструкторской работы по созданию в АО "ОКБ "Электроавтоматика" специализированного программного обеспечения одного из видов обеспечения. Такое программное обеспечение направлено на поддержку САПР для автоматизированного формирования взаимосвязанных текстовых конструкторских документов на программно-управляемые изделия и контроля жизненного цикла документов и содержащихся в них инженерных данных.

Описание жизненного цикла документации в разработанной системе проектирования

Для решения задачи автоматизированного формирования взаимосвязанных текстовых конструкторских документов на программно-управляемые изделия была использована среда Microsoft Visual Studio. Программное средство написано на языке C# с использованием WPF (Windows Presentation Foundation). В состав предлагаемого решения входит система контроля версий технической документации, которая осуществляет постоянный контроль версий каждого документа в ходе его жизненного цикла, представленного на рис.1.

При создании каждого нового технического документа статус документа автоматически переходит в состояние "черновик" (*draft*) с версией 1.0. Это состояние используется для редактирования документа. По окончании редактирования документа версия документа подвергается формальной инспекции, для чего состояние документа переводится в статус "проверка" (*proposed*). С этого момента редактирование документа запрещается. По результатам проверки версия документа либо становится новой базовой версией единицы конфигурации (ЕК), переходя в состояние "утверждено" (*approved*), либо документ отправляется на доработку разработчику, переходя в состояние статуса документа "отклонен" (*declined*). В обоих случаях версия документа фиксируется в архивной системе предприятия и ее дальнейшее редактирование не допускается. При необходимости внести изменения в документ выполняется генерация следующей версии этого документа. При этом в зависимости от состояния текущей версии документа (*approved* или *declined*) выполняется присвоение нового номера версии для генерируемой версии документа.

Все версии документа сохраняются в электронном архиве предприятия. В предлагаемой системе проектирования обеспечена возможность просмотра любой выбранной пользователем версии документа, а также возможность сравнения двух любых версий одного и того же документа.

Группировка нескольких версий различных документов, связанных между собой в едином разрабатываемом проекте, осуществляется с помощью конфигураций документов. Конфигурация документов может содержать те версии каждого из документов проекта, которые вошли в релиз документации на разработку. Другим примером конфигурации документов является сочетание версий различных документов на момент завершения определенной стадии жизненного цикла проекта, предполагающего разработку технической документации на изделие. Конфигурация документов не имеет версий и является особой формой ЕК с упрощенным числом воз-

Рис. 1. Обобщенный жизненный цикл технической документации

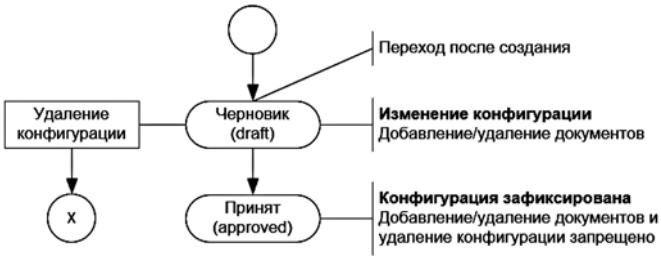

Рис. 2. Состояния конфигурации документов

можных состояний: черновик (*draft*) и утверждено (*approved*).

Возможные состояния конфигурации документов и переходы между ними представлены на рис. 2.

После перехода конфигурации документов в состояние *approved* ее изменение (добавление и удаление документов, изменение имени конфигурации документов) и удаление самой конфигурации документов запрещено.

Описание инструментального средства контроля версий технических документов

В терминах объектно-ориентированного подхода к разработке документации каждому объекту поставлен в соответствие класс, объекты которого будут использоваться для хранения сущностей. На рис. 3 показан фрагмент разработанной формальной информационной модели в виде UML-диаграммы классов, иллюстрирующей состав и иерархию объектов проектирования, их основные атрибуты и операции в контексте управления жизненным циклом технической документации.

Интерфейс разработанной программной среды контроля версий документов имеет древовидную структуру, он представлен на рис. 4.

В инструментальном средстве реализована возможность просмотра истории изменений документа от версии к версии. История изменений документов представляет собой документ в формате TXT. Он включает в себя:

- стоки документа, содержащие имя измененного элемента;
- уникальный идентификатор документа, присвоенный при создании и не изменяющийся на протяжении всего жизненного цикла;

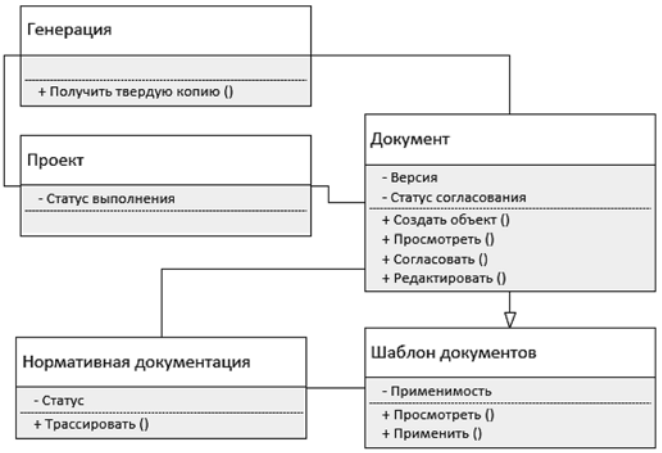

Рис. 3. UML-диаграмма классов объектов проектирования

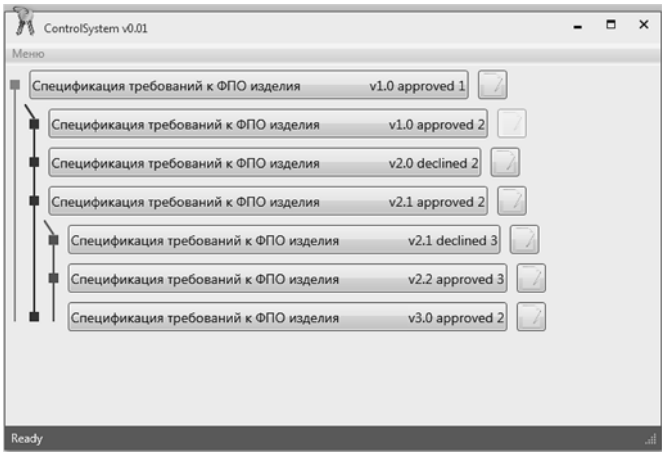

Рис. 4. Интерфейс программной среды контроля версий документов: ФПО — функциональное программное обеспечение

- тип изменения документа;
- предыдущее и новое значения измененных в документе инженерных данных.

Идентификатор представляет собой сгенерированный статистически уникальный 128-битный идентификатор GUID (*Globally Unique Identifier*). Пример такого документа представлен на рис. 5.

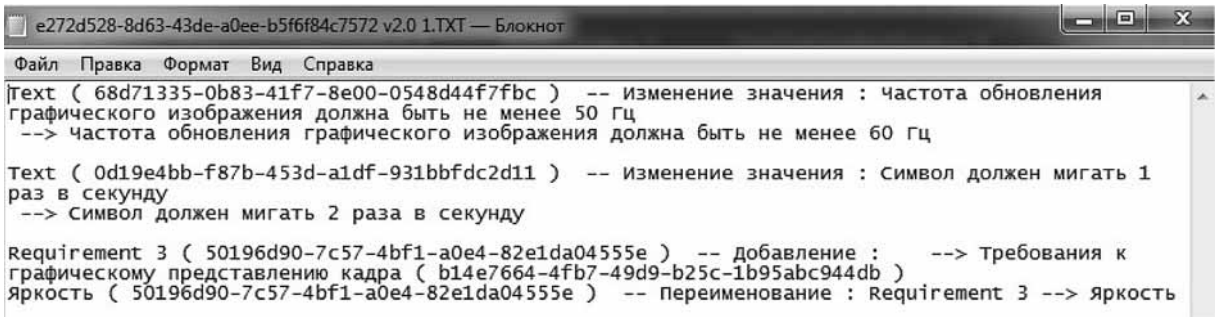

Рис. 5. История изменения документа

Результаты экспериментов

Для исследования показателей качества предложенного программного решения была проведена серия экспериментов. Показателем качества программного средства было выбрано время работы программы по подготовке и сохранению на носителе памяти ЭВМ текстовых конструкторских документов в разных форматах. В техническом задании было сформулировано требование к времени сохранения проекта в разрабатываемом программном обеспечении — не более 1 мин. Программная поддержка проведения эксперимента выполнена с использованием библиотеки Microsoft.Office.Interop.Word.dll версии 14.0, а также с использованием языка C#. Тестирование времени сохранения документа на носителе памяти проводилось замером времени выполнения следующего кода:

```
Microsoft.Office.Interop.Word._Application appWord= new
Microsoft.Office.Interop.Word.Application();
// Заполнение документа контентом
CreateAndFillActiveDoc(appWord);
object fileName;
object FileFormat =
Word.WdSaveFormat.wdFormatRTF;
    appWord.ActiveDocument.SaveAs(ref fileName, ref FileFormat);
    appWord.Quit();
```

Расширение сохраняемого документа задается переменной *FileFormat*. В приведенном выше примере выбран формат *RTF*. Были также исследованы форматы *PDF, XPS, DOC, DOCX, ODT, MHTML*. Сохранялись данные объемом три страницы формата А4, содержащие таблицу, изображение и текст. Результаты эксперимента представлены на рис. 6 в виде столбиковой диаграммы.

Исходя из полученных результатов сделан вывод, что сохранять промежуточные версии каждого технического документа проекта в архивной системе промышленного предприятия целесообразно в формате RTF, что позволяет снизить время разработчика при сохранении большого объема инженерных данных проекта. Так как средний объем комплекта технической документации на программно-управляемое устройство не превышает 1000 страниц, то можно сделать вывод, что на его сохранение в формате RTF понадобится не более минуты. Следовательно, разработанное программное обеспечение удовлетворяет вышеуказанным требованиям.

Заключение

Описанное программное обеспечение автоматизирует подготовку конструкторской и программной документации на приборостроительном предприятии при разработке программно-управляемых изделий, а также минимизирует число ошибок при разработке этих документов, контролируя жизненный цикл инженерных данных. Как следствие, уменьшаются трудозатраты на разработку комплектов документов, что делает создание системы автоматизации оформления технической документации экономически выгодной.

Описана новая подсистема контроля версий, которая обеспечивает постоянный контроль жизненно-

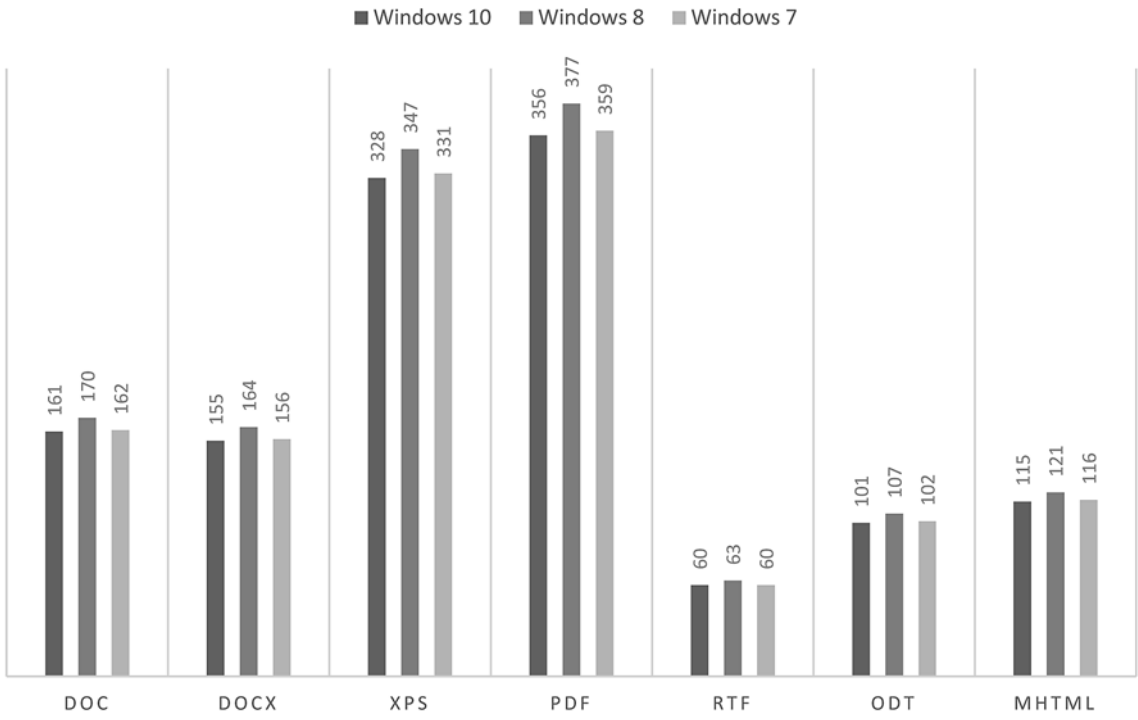

Рис. 6. Время сохранения данных при различных форматах текстового документа, с

го цикла документов и инженерных данных, в целях повышения качества разработки документов. Пользователь программного обеспечения имеет возможность оценить этапы жизненного цикла документа, кем и когда были внесены определенные изменения в документ, а также вернуться к предыдущей версии документа в случае необходимости. Представлены результаты эксперимента по измерению времени сохранения документов в популярных текстовых форматах.

Апробация разработанного программного обеспечения была реализована в АО "ОКБ "Электроавтоматика" (г. Санкт-Петербург) с использованием действующих в области авиационного приборостроения отечественных стандартов и стандартов ARINC (Aeronautical Radio Inc., США), в частности, ARINC 651-655, при подготовке документации на бортовые цифровые вычислительные системы.

Описанное программное обеспечение разработано на языке программирования C# с использованием WPF и функционирует на базе инструментальной ЭВМ со следующими характеристиками: MSI GP72 7RD-254RU, процессор Intel(R) Core(TM) i5-7300HQ, 4 ядра, тактовая частота 3,5 ГГц, оперативная память 8 Gb под управлением операционной системы Windows 10.

Список литературы

1. **Авдеева М., Чиркин А.** Перевод бумажной документации в электронный вид // САПР и графика. 2004. № 1. С. 70—72.
2. **Садовников Д., Ноздрин А., Ширяев Н.** Система управления технической и проектно-конструкторской документацией // САПР и графика. 2002. № 5. С. 74—77.
3. **Кукаренко Е., Молочко Д.** Управление потоками знаний в техническом документообороте предприятия // САПР и графика. 2001. № 10. С. 35—37.
4. **Бычков И., Ващук Ю.** Конструкторская спецификация — информационная основа управления предприятием // САПР и графика. 2001. № 9. С. 90—95.
5. **ГОСТ 2.105—95** Единая система конструкторской документации (ЕСКД). Общие требования к текстовым документам (с Изменением N 1). М.: Стандартинформ, 2011.
6. **ГОСТ 2.104—68** Единая система конструкторской документации (ЕСКД). Основные надписи (с Изменениями N 1—7). М.: ИПК Издательство стандартов, 2002.
7. **Брахутин А. Г.** CALS выходит на федеральный уровень // Вестник авиации и космонавтики. 2001. № 5. С.26-27.
8. **Гатчин Ю. А., Жаринов И. О., Жаринов О. О.** Архитектура программного обеспечения автоматизированного рабочего места разработчика бортового авиационного оборудования // Научно-технический вестник информационных технологий, механики и оптики. 2012. № 2. С. 140—141.
9. **Гатчин И. Ю., Жаринов И. О., Жаринов О. О., Косенков П. А.** Реализация жизненного цикла "проектирование—производство—эксплуатация" бортового оборудования на предприятиях авиационной промышленности // Научно-технический вестник информационных технологий, механики и оптики. 2012. № 2. С. 141—143.
10. **Парамонов П. П., Гатчин Ю. А., Жаринов И. О., Жаринов О. О., Дейко М. С.** Принципы построения отраслевой системы автоматизированного проектирования в авиационном приборостроении // Научно-технический вестник информационных технологий, механики и оптики. 2012. № 6. С. 111—117.
11. **Utkin S. B., Batova S. V., Blagonravov S. A., Konovalov P. V., Zharinov I. O.** Automated construction of software configuration tables for real-time systems in avionics // Programming and Computer Software. 2015. Vol. 41, No. 4. P. 219—223.
12. **Благонравов С. А., Уткин С. Б., Батова С. В., Коновалов П. В.** Опыт применения технологии эмуляции процессов при разработке компонентов программного обеспечения авиационных систем // Программная инженерия. 2015. № 8. С. 18—25.
13. **Шек-Иовсепянц Р. А., Жаринов И. О.** Генерация проектных решений бортового оборудования с использованием аппарата генетических алгоритмов // Научно-технический вестник информационных технологий, механики и оптики. 2010. № 3. С. 67—70.
14. **Жаринов И. О., Жаринов О. О., Шек-Иовсепянц Р. А., Суслов В. Д.** Оценка снижения трудоемкости подготовки конструкторской документации с использованием CALS-технологии в приборостроении // Научно-технический вестник информационных технологий, механики и оптики, 2012. № 4. С. 151—153.

The Software for Automation Monitoring of Life Cycle of Text Documentation on Program-Driven Products

S. A. Leonovets[1, 2], ser2694@ya.ru, **A. V. Gurjanov**[2], postmaster@elavt.spb.ru, **A. V. Shukalov**[1, 2], aviation78@mail.ru, **I. O. Zharinov**[1, 2], igor_rabota@pisem.net

[1]Saint Petersburg National Research University of Information Technologies, Mechanics and Optics (ITMO University), Saint Petersburg, 197101, Russian Federation,

[2]Design Bureau "Electroavtomatika", Saint Petersburg, 198095, Russian Federation

Corresponding author:

Zharinov Igor O., Chef of Department, Saint Petersburg National Research University of Information Technologies, Mechanics and Optics (ITMO University), 197101, Saint Petersburg, Russian Federation
e-mail: igor_rabota@pisem.net

Received on December 12, 2017
Accepted on December 27, 2017

The task of automating of process of preparation, storage and monitoring of versions of text designer and program documentation by means of the specialized software is considered. Automation the preparation of documentation is based on processing of engineering data which is contained in specifications and technical documentation. Data handling assumes existence of strictly structured electronic documents prepared in widespread formats according to templates on the basis of industry standards and generating the text techincal documenation by means of an automated method. Further life cycle of the document and the technical data which are contained in it is controlled. At each stage of life cycle the archive data storage is executed. The article presents the results of a research on estimating the performance of the text editor when using various widespread document formats for the stages of automated monitoring and storage of them.The new developed software and work benches facilitating development of the instrumental equipment is described.

Keywords: *design documentation, automation, text document, life cycle of the document*

For citation:

Leonovets S. A., Gurjanov A. V., Shukalov A. V., Zharinov I. O. The Software for Automation Monitoring of Life Cycle of Text Documentation on Program-Driven Products, *Programmnaya ingeneria,* 2018, vol. 9, no. 2, pp. 91—96.

DOI: 10.17587/prin.9.91-96

References

1. **Avdeeva M., Chirkin A.** Perevod bumazhnoj dokumentacii v jelektronnyj vid (Convert paper documents into electronic form), *SAPR i grafika*, 2004, no. 1, pp. 70—72 (in Russian).

2. **Sadovnikov D., Nozdrin A., Shirjaev N.** Sistema upravlenija tehnicheskoj i proektno-konstruktorskoj dokumentaciej (The control system of technical and design documentation), *SAPR i grafika*, 2002, no. 5, pp. 74—77 (in Russian).

3. **Kukarenko E., Molochko D.** Upravlenie potokami znanij v tehnicheskom dokumentooborote predprijatija (Flow management knowledge in the technical document of the enterprise), *SAPR i grafika*, 2001, no. 10, pp. 35—37 (In Russian).

4. **Bychkov I., Vashhuk Ju.** Konstruktorskaja specifikacija — informacionnaja osnova upravlenija predprijatiem (The design specification — the basis of enterprise management information), *SAPR i grafika*, 2001, no. 9, pp. 90—95 (in Russian).

5. **GOST** 2.105—95 ESKD. General requirements for text documents (with Change No. 1). Moscow, Standartinform, 2011 (in Russian).

6. **GOST** 2.104—68. Software Unified system of design documentation (ESKD). The main inscriptions (with Changes N 1—7). Moscow, IPK Publishing House of Standards, 2002 (in Russian).

7. **Brahutin A. G.** CALS vyhodit na federal'nyj uroven' (CALS goes to the federal level), *Vestnik aviacii i kosmonavtiki*, 2001, no. 5, pp. 26—27 (in Russian).

8. **Gatchin Yu. A., Zharinov I. O., Zharinov O. O.** Software Architecture for the Automated Workplace of the Onboard Aviation Equipment Developer, *Nauchno-tekhnicheskii vestnik informatsionnykh tekhnologii, mekhaniki i optiki*, 2012, no. 2, pp. 140—141 (in Russian).

9. **Gatchin I. Yu., Zharinov I. O., Zharinov O. O., Kosenkov P. A.** Realizacija zhiznennogo cikla "proektirovanie—proizvodstvo—jeksploatacija" bortovogo oborudovanija na predprijatijah aviacionnoj promyshlennosti (The implementation of the life cycle "design-production-operate" on-board equipment to the aviation industry enterprises), *Nauchno-tekhnicheskii vestnik informatsionnykh tekhnologii, mekhaniki i optiki*, 2012, no. 2, pp. 141—143 (in Russian).

10. **Paramonov P. P., Gatchin Yu. A., Zharinov I. O., Zharinov O. O., Deiko M. S.** Principy postroenija otraslevoj sistemy avtomatizirovannogo proektirovanija v aviacionnom priborostroenii (Principles of Branch System Creation for the Automated Design in Aviation Instrumentation), *Nauchno-tekhnicheskii vestnik informatsionnykh tekhnologii, mekhaniki i optiki*, 2012, no. 6, pp. 111—117 (in Russian).

11. **Utkin S. B., Batova S. V., Blagonravov S. A., Konovalov P. V., Zharinov I. O.** Automated construction of software configuration tables for real-time systems in avionics. *Programming and Computer Software*, 2015, vol. 41, no. 4, pp. 219—223.

12. **Blagonravov S. A., Utkin S. B., Batova S. V., Konovalov P. V.** Opyt primenenija tehnologii jemuljacii processov pri razrabotke kompo-nentov programmnogo obespechenija aviacionnyh sistem (Using Emulation Technics in the Development of Avionics Software Components), *Programmnaya ingeneria*, 2015, no. 8, pp. 18—25 (in Russian).

13. **Shek-Iovsepjanc R. A., Zharinov I. O.** Generacija proektnyh reshenij bortovogo oborudovanija s ispol'zovaniem apparata geneticheskih algoritmov (Design Generation of the Avionic Equipment by Genetic Algorithms), *Nauchno-tekhnicheskii vestnik informatsionnykh tekhnologii, mekhaniki i optiki*, 2010, no. 3, pp. 67—70 (in Russian).

14. **Zharinov I. O., Zharinov O. O., Shek-Iovsepjanc R. A., Suslov V. D.** Ocenka snizhenija trudoemkosti podgotovki konstruktorskoj dokumen-tacii s ispol'zovaniem CALS-tehnologii v priborostroenii (Assessment of reducing the complexity of the preparation of the design documentation using the CALS-technologies in instrument), *Nauchno-tekhnicheskii vestnik informatsionnykh tekhnologii, mekhaniki i optiki*, 2012, no. 4, pp. 151—153 (in Russian).

ООО "Издательство "Новые технологии". 107076, Москва, Стромынский пер., 4
Технический редактор *Е. М. Патрушева.* Корректор *Е. В. Комиссарова*

Сдано в набор 12.12.2017 г. Подписано в печать 23.01.2018 г. Формат 60×88 1/8. Заказ PI218
Цена свободная.

Оригинал-макет ООО "Адвансед солюшнз". Отпечатано в ООО "Адвансед солюшнз".
119071, г. Москва, Ленинский пр-т, д. 19, стр. 1. Сайт: **www.aov.ru**

Рисунки к статье А. Бернадотт
«АНАЛИЗ НАУЧНОГО ТЕКСТА И НОВЫЕ МИРОВЫЕ ТЕНДЕНЦИИ»

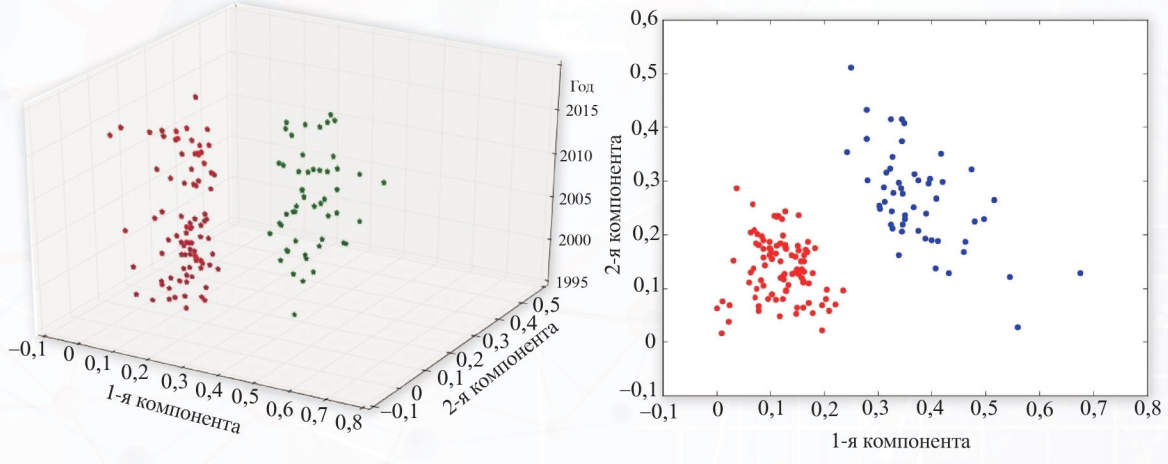

Рис. 3. Метод *k*-средних. Кластер, соответствующий «фундаментальным» («теоретическим») документам, отмечен красным. Синим и зеленым отмечен кластер, соотвествующий статьям «практического» («прикладного») содержания. Каждой точке соответствует 50 статей

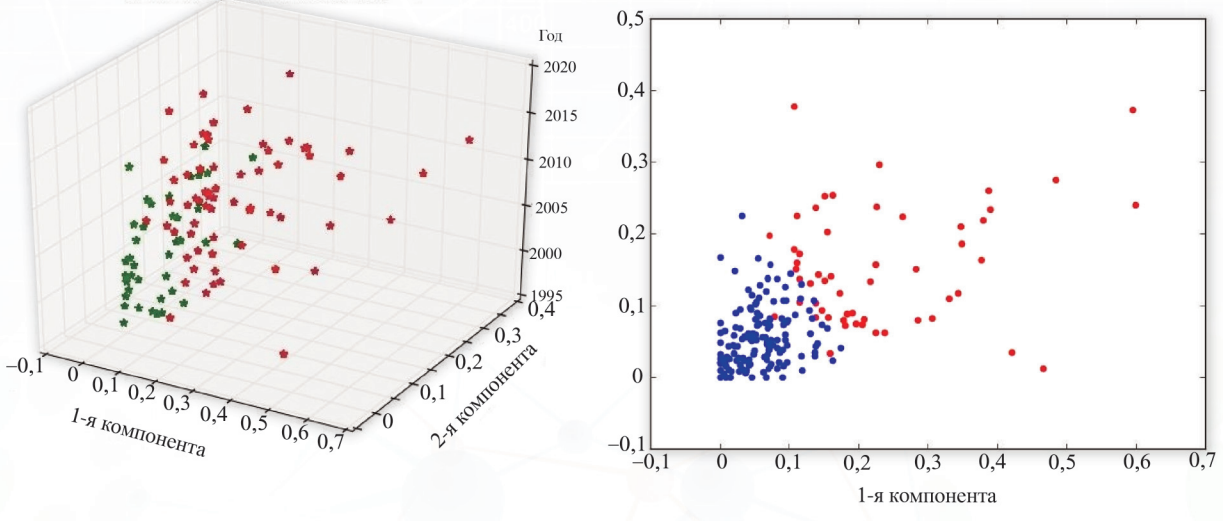

Рис. 4. Метод *k*-средних. Кластер, соответствующий документам с превалированием слов из «коммерческого» слово-класса, отмечен красным.Синим и зеленым отмечен кластер, соответствующий статьям некоммерческой направленности. Каждой точке соответствует 50 статей

Издательство «НОВЫЕ ТЕХНОЛОГИИ»

выпускает *научно-технические журналы*

Теоретический и прикладной научно-технический журнал

ПРОГРАММНАЯ ИНЖЕНЕРИЯ

В журнале освещаются состояние и тенденции развития основных направлений индустрии программного обеспечения, связанных с проектированием, конструированием, архитектурой, обеспечением качества и сопровождением жизненного цикла программного обеспечения, а также рассматриваются достижения в области создания и эксплуатации прикладных программно-информационных систем во всех областях челове ческой деятельности.

Подписные индексы по каталогам:
«Роспечать» – 22765; «Пресса России» – 39795

Ежемесячный теоретический и прикладной научно-технический журнал

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

В журнале освещаются современное состояние, тенденции и перспективы развития основных направлений в области разработки, производства и применения информационных технологий.

Подписные индексы по каталогам:
«Роспечать» – 72656;
«Пресса России» – 94033

Ежемесячный междисциплинарный теоретический и прикладной научно-технический журнал

НАНО- и МИКРОСИСТЕМНАЯ ТЕХНИКА

В журнале освещаются современное состояние, тенденции и перспективы развития нано- и микросистемной техники, рассматриваются вопрос разработки и внедрения нанои микросистем в различные области науки, технологии и производства.

Подписные индексы по каталогам:
«Роспечать» – 79493;
«Пресса России» – 27849

Ежемесячный теоретический и прикладной научно-технический журнал

МЕХАТРОНИКА, АВТОМАТИЗАЦИЯ, УПРАВЛЕНИЕ

В журнале освещаются достижения в области мехатроники, интегрирующей механику, электронику, автоматику и информатику в целях совершенствования технологий производства и создания техники новых поколений. Рассматриваются актуальные проблемы теории и практики автомати ческого и автоматизированного управления техническими объектами и технологическими процессами в промышленности, энергетике и на транспорте.

Подписные индексы по каталогам:
«Роспечать» – 79492;
«Пресса России» – 27848

Научно-практический и учебно-методический журнал

БЕЗОПАСНОСТЬ ЖИЗНЕДЕЯТЕЛЬНОСТИ

В журнале освещаются достижения и перспективы в области исследований, обеспечения и совершенствования защиты человека от всех видов опасностей производственной и природной среды, их контроля, мониторинга, предотвращения, ликвидации последствий аварий и катастроф, образования в сфере безопасности жизнедеятельности.

Подписные индексы по каталогам:
«Роспечать» – 79963;
«Пресса России» – 94032

Адрес редакции журналов для авторов и подписчиков:
107076, Москва, Стромынский пер., 4. Издательство "НОВЫЕ ТЕХНОЛОГИИ".
Тел.: (499) 269-55-10, 269-53-97. Факс: (499) 269-55-10. E-mail: antonov@novtex.ru

ISSN 2220-3397. Программная инженерия. 2018. Том 9. № 2. 49–96. Индексы: «Роспечать» – 22765, «Пресса России» – 39795