На правах рукописи

Иванов Алексей Владимирович

АРХИТЕКТУРА И ПРОГРАММНАЯ ИНФРАСТРУКТУРА СИСТЕМ УПРАВЛЕНИЯ КОНТЕНТОМ И МОДЕЛИ ОПИСАНИЯ ИХ ФУНКЦИОНИРОВАНИЯ

05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Москва – 2018

Работа выполнена в Федеральном исследовательском центре «Информатика и управление» Российской академии наук

Научный руководитель: **Босов Алексей Вячеславович**
доктор технических наук, доцент, главный научный сотрудник Федерального исследовательского центра «Информатика и управление» Российской академии наук

Официальные оппоненты: **Наумов Андрей Викторович**
доктор физико-математических наук, доцент, профессор кафедры «Теория вероятностей и компьютерное моделирование» факультета №8 «Информационные технологии и прикладная математика» Московского авиационного института (национального исследовательского университета)

Выголов Олег Вячеславович
кандидат технических наук, начальник лаборатории систем улучшенного и синтезированного видения Государственного научно-исследовательского института авиационных систем

Ведущая организация: Федеральное государственное бюджетное учреждение науки Институт проблем передачи информации им. А.А. Харкевича Российской академии наук

Защита состоится «27» июня 2018 года в 15 ч. 00 мин. на заседании диссертационного совета Д 002.073.02 при Федеральном исследовательском центре «Информатика и управление» Российской академии наук по адресу 119333, Москва, ул. Вавилова, д. 44, к. 2.

С диссертацией можно ознакомиться в библиотеке Федерального исследовательского центра «Информатика и управление» Российской академии наук по адресу г. Москва, ул. Вавилова, д.44, к. 2 и на сайте www.frccsc.ru.

Автореферат разослан « » 2018 г.

Ученый секретарь
диссертационного совета Р. В. Разумчик

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. Система управления контентом (Content Management System, CMS) – компонент многих информационных систем: интернет и интранет порталов, web-сайтов и цифровых библиотек. CMS применяется для подготовки, публикации и коллективной работы с информацией, представленной в различных формах (контентом).

CMS является ключевым компонентом таких решений распределенной обработки данных, как порталы. Портал – универсальное средство, технология доступа к распределенным системам, в частности к информации, как из внешних источников, так и из собственного хранилища[1]. Задачи отображения информации, а также управления собственным хранилищем портала как раз и требуют наличия CMS.

Сложность построения интегрированных информационных систем во многом связана с организацией эффективного взаимодействия их компонентов. В случае CMS какие-либо стандарты по организации их взаимодействия с другими системами долгое время отсутствовали, а архитектуры большинства систем были закрытыми.

Хотя в последние годы выросло число систем с открытой архитектурой, появился ряд стандартов, регламентирующих обмен контентом и взаимодействие с CMS и её компонентами, пока нельзя признать эту работу достаточной, а поддержку стандартов – повсеместной.

Архитектурные ограничения доступных CMS делают оправданной, актуальной, а зачастую и единственно возможной разработку собственных систем, адаптированных под задачи интеграции.

Так, например, для управления хранилищем портала Российской академии наук http://www.ras.ru, содержащего около 40000 документов и информацию о 20000 персон и 2100 организаций, пригодны многие коммерческие CMS. Однако, помимо собственного хранилища, портал предоставляет доступ к 6 базам данных материалов архива РАН, мультимедийному архиву, архиву журнала «Вестник РАН», серии электронных коллекций по научному наследию и другим ресурсам. Также в портал интегрирован ряд информационных систем Президиума РАН. Организация управления такими разнообразными ресурсами во многих CMS осложняется их внутренними ограничениями и требует значительных усилий по разработке с участием производителей CMS. Такой сценарий по затратам мало отличим от создания собственной системы.

Таким образом, задачи анализа и проектирования собственной CMS, ориентированной на применение в составе интеграционного решения, являются весьма *актуальными*.

---

[1] Черняк Л. Корпоративный портал // PC Week/RE URL: http://www.pcweek.ru/idea/article/detail.php?ID=51690 (дата обращения 18.04.2016)

В первых двух частях диссертационной работы изложено решение указанных задач (отображения и управления контентом) применительно к CMS для Информационного web-портала – разработки, выполненной в ФИЦ ИУ РАН[2], и используемой в качестве интеграционной платформы в ряде проектов. Также в них отражено решение задачи сопровождения конечного программного средства, в данном случае заключающейся в обеспечении эффективного управления постоянно растущим массивом контента, расширением состава различных внешних информационных источников, а также организации доступа к ним.

Процесс разработки, помимо анализа требований, определения подлежащего реализации функционала, выбора программной и технической платформ, включает организацию тестирования CMS, необходимого для оценки ее работоспособности при расчетных нагрузках и выявления программных ошибок.

Обычно организация тестовой нагрузки сводится к созданию потока запросов, соответствующего расчетному или стрессовому уровню. Такую нагрузку сложно признать адекватной, поскольку она не отражает реальную активность пользователей и не позволяет достоверно определить реакцию продукта на эту активность.

Хотя число пользователей и поддается оценке, трудность представляет определение их ожидаемой активности, т.е. частоты обращений к компонентам системы, смены уровней активности, различия этих уровней и характера смены одного уровня активности другим.

Трудность определения реакции системы обусловлена наличием групп пользователей, использующих различный функционал системы и создающих на нее различную нагрузку, и непредсказуемым взаимным влиянием других приложений, использующих общие с CMS аппаратные и телекоммуникационные ресурсы.

Компании уровня Google обладают значительными накопленными шаблонами нагрузок, применяемыми при тестировании новых разработок[3], однако рядовые разработчики такой возможности лишены.

В диссертационной работе предлагается рассматривать пользовательскую активность как носящую случайный характер. Традиционно применяемые линейные регрессионные модели не позволяют описать ее удовлетворительно. В связи с этим для моделирования активности пользователей предлагается использовать аппарат стохастических ди-

---

[2] Свидетельство о регистрации программы для ЭВМ №2005612992 от 20 сентября 2005г.

[3] Schwarzkopf M., Konwinskiz A., Abd-El-Malek M., Wilkes J. Omega: flexible, scalable schedulers for large compute clusters // URL: http://eurosys2013.tudos.org/wp-content/uploads/2013/paper/Schwarzkopf.pdf (дата обращения 18.04.2014)

намических систем. Для исследования реальной пользовательской активности средства сбора характеризующих ее данных были включены в состав рассматриваемой в работе CMS. Экспертный анализ полученных данных позволил предложить модель пользовательской активности в форме стохастической динамической системы с дискретным временем. Изначально заложенное в модель предположение о наличии типичных состояний пользовательской активности позволило достаточно детально управлять параметрами модели и обеспечить адекватность моделируемой активности реально наблюдаемым данным.

Предложенная модель активности пользователей была применена на практике для тестирования CMS. Используя модель, отдельное приложение формировало поток запросов. Запросы выбирались из блоков типовых запросов, ранее вводившихся пользователями при работе с системой. Протоколы работы системы использовались для диагностики компонентов и контроля расходования ресурсов.

Следует отметить, что такой поход к тестированию является *новым,* стохастические модели для формирования тестовой нагрузки сколь-нибудь существенно не применялись. Более того, подход оказывается эффективным (что обосновывает перспективность как его практического, так и теоретического развития) и в других задачах. Так, в рамках диссертации рассмотрена еще одна задача, связанная с анализом состояния внешних источников информации портала.

Данные о времени выполнения запросов внешними источниками информации для CMS не известны, и могут существенно меняться даже для идентичных запросов. На время выполнения запросов влияют нагрузка аппаратных и телекоммуникационных ресурсов, наличие данных в кэше, а также семантика запроса. Использование внешними источниками общих аппаратных ресурсов приводит к взаимному влиянию на время выполнения запросов. Наибольший же вклад в неопределенность вносят действия пользователей, поскольку заранее не известны ни состав требующихся им источников информации, ни интенсивность и структура формируемых ими запросов.

Точное вычисление времени выполнения запросов по-видимому нереализуемо. Однако возможна оценка времени их выполнения, которая может использоваться для планирования выполнения запросов и повышения эффективности использования ресурсов системы.

В диссертационной работе на основе экспертного анализа массива экспериментальных данных по временам выполнения запросов предложена модель показателя эффективности источника в форме стохастической динамической системы с дискретным временем.

Целью работы являются разработка концепции CMS для интеграционной программной платформы, предоставляемой Информационным web-порталом и описание среды функционирования этой си-

стемы адекватными математическими моделями. Требующиеся для достижения цели исследования задачи:

1) анализ и сравнение существующих подходов к реализации систем управления web-контентом, формирование требований к такой системе, входящей в состав информационного web-портала;

2) разработка архитектуры и программной реализации компонентов системы, их реализация, получение экспериментальных данных;

3) формирование универсального подхода к описанию состояния среды функционирования CMS, на основе которого возможно предложить адекватные математические модели;

4) проверка адекватности предложенных моделей полученным данным и анализ применимости моделей в реальных задачах (в частности, задачах тестирования и оптимизации функционирования).

Объектом исследования диссертации являются модели для анализа функционирования программных систем (в частности CMS) под нагрузкой, имитирующей пользовательскую активность, процессы и алгоритмы проектирования, применявшиеся в ходе создания собственного программного решения. Для анализа предложено использовать модели на основе стохастических динамических систем. Данные модели служат основой для решения задач повышения эффективности и надежности процессов обработки данных в подсистемах Информационного web-портала, включая моделирование и организацию тестовой нагрузки.

Разработанное программное обеспечение имеет самостоятельное практическое значение и было использовано в целом ряде информационных систем.

Предмет исследования составляют алгоритмы и программная инфраструктура для организации распределенной обработки данных в портальной CMS и информационные процессы, влияющие на ее функционирование.

Результаты, выносимые на зашиту:

1. Проектные решения по архитектуре и программной инфраструктуре портальной CMS, полученные с использованием методов объектно-ориентированного проектирования и анализа алгоритмов и программ.

2. Программа для ЭВМ, обеспечивающая управление контентом и предназначенная для использования в составе интеграционной платформы информационного web-портала.

3. Математическая модель динамики показателя пользовательской активности, представленная стохастической динамической системой, использующей авторегрессионную модель с переключениями.

4. Математическая модель динамики показателя эффективности информационного источника, представленная стохастической динами-

ческой системой наблюдений, использующей авторегрессионную модель с переключениями. Предложен подход к определению параметров модели на основе простейшего статистического анализа среды функционирования web-портала.

5. Алгоритмические и программные решения для нагрузочного тестирования и анализа вероятностно-временных характеристик программной системы на основе интеграционной платформы Информационного web-портала.

Научная новизна. В работе автором самостоятельно получены следующие новые теоретические и научно-технические результаты:

– выполнены проект и разработка самостоятельной CMS;

– предложен, исследован и реализован подход к моделированию и организации тестовой нагрузки;

– предложен и исследован подход к моделированию состояния информационных источников.

Методы исследования. Методическую основу исследования в части задач моделирования (среды функционирования CMS и показателя эффективности информационного источника) обеспечивает теория стохастических динамических систем. Анализ массивов данных, полученных в ходе эксплуатации системы, проведен методами математической статистики.

Для решения поставленных задач в диссертации использовались методы объектно-ориентированного проектирования и анализа алгоритмов и программ и принципы построения сервисно-ориентированной архитектуры (Service OA).

В основе проектных решений лежат методы объектно-ориентированного программирования, компонентный подход к разработке программ, а также стандарты и спецификации Open Management Group (UML, PKI), World Wide Web Consortium (HTML, CSS, XML, WSDL, SOAP), European Computer Manufacturers Association (JavaScript), Internet Engineering Task Force (RFC).

Практическая значимость полученных результатов подтверждается рядом выполненных проектов – внедренных в практическую деятельность информационных систем, в которых была применена созданная CMS. Результаты диссертации применены в программе Президиума РАН «Информатизация» (2001-2010 гг.), учебном портале, а также серии проектов, выполняемых ФИЦ ИУ РАН в рамках работ по созданию систем специального назначения.

Действующие реализации программного решения, а также существенные объемы статистических данных, полученных в ходе организации нагрузочного тестирования, подтверждают **достоверность результатов** полученных в диссертации.

Апробация работы. Основные положения и результаты диссертационной работы докладывались на конференциях и семинарах:

– «Информационное обеспечение науки», Таруса, 2003;

– II научная сессия ИПИ РАН, 2005;

– восьмая Всероссийская научная конференция «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», Суздаль, 17-19 октября 2006;

– XII Российская конференция с международным участием «Распределенные информационно-вычислительные ресурсы» (DICR2008), Академгородок, Новосибирск, 5-7 ноября 2008;

– XI Всероссийский симпозиум по прикладной и промышленной математике (весенняя сессия), Кисловодск, 1-8 мая 2010.

Кроме того, подходы, отраженные в работе, неоднократно представлялись на научных семинарах в ФИЦ ИУ РАН, ВЦ РАН, МСЦ РАН, ИСА РАН.

Публикации. По теме диссертации автором опубликовано 13 работ; 8 из них опубликованы в рецензируемых научных изданиях, рекомендованных ВАК [1–8]. Получено одно свидетельство о регистрации программы для ЭВМ.

Структура и объем диссертации. Диссертация состоит из введения, четырех глав, заключения и списка использованных источников (111 наименований). Работа изложена на 160 страницах, содержит 34 рисунка и 10 таблиц.

СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность работы, дана краткая характеристика тематических публикаций и описана общая структура диссертации.

Первая глава посвящена истории развития, классификации, анализу функциональности и принципов проектирования систем управления контентом. Приведены основные понятия предметной области, рассмотрены возможности ряда доступных CMS. Выделены тенденции развития этих систем, такие как применение компонентного подхода к управлению контентом, ведущийся процесс стандартизации, наличие взаимного влияния различных категорий продуктов.

Результатом проделанного анализа является выбор и обоснование подходов к проектированию собственной CMS. Использование CMS в составе Информационного web-портала предполагает ее взаимодействие с внешними информационными источниками, включая не только доступ к их данным, но и поддержку полнотекстового и атрибутного поиска, возможность настройки системы на публикацию различных типов данных и др. Такой функционал требует от CMS поддержки

возможности работы со средствами интеграции данных, и наличия архитектуры позволяющей расширять функциональность системы.

На рынке решений для управления контентом можно выделить две категории продуктов: для управления web и корпоративным контентом. Для первой характерна невысокая стоимость, но поддержка работы с внешними информационными источниками ограничена или рудиментарна. Вторая изначально пригодна для задач интеграции, но обладает высокой стоимостью и сложностью внедрения, требующей взаимодействия с производителем системы. Для обеих категорий характерны следующие ограничения:

– закрытость архитектуры, затрудняющая их интеграцию с другими системами;

– недостаточная стандартизация форматов данных и процессов управления контентом.

В связи с этим, при проектировании собственного решения была сделана ставка на открытость архитектуры, и максимальное использование имеющихся стандартов, протоколов и технологий Интернет.

Кроме того, идеология разработки CMS для Информационного web-портала основана на более широком, по сравнению с традиционным, взгляде на CMS, как на систему, предназначенную для управления практически любыми типами содержания, предоставляемыми разнородными источниками информации. Естественно, что такой взгляд предполагает наличие согласованной архитектуры всех служб и подсистем портала, а не только CMS.

Во второй главе рассматриваются технические требования, архитектурные решения и практическая реализация системы управления контентом для Информационного web-портала.

К основным техническим требованиям, которым должно удовлетворять создаваемое программное решение, отнесены:

– возможность управления жизненным циклом различных видов контента, включающих структурированную, слабоструктурированную и неструктурированную информацию;

– возможность использования для редактирования контента как приложений, написанных по классической технологии «толстого» клиента, так и web-интерфейс;

– наличие средств автоматизации управления жизненным циклом контента, включая публикацию контента по расписанию и автоматическую архивацию контента;

– поддержка средств интеграции контента из разных информационных источников и формирования контента как сочетания статической (редко меняющейся) и динамической (генерируемой при обращении к информационным источникам и службам) информации;

– реализация программно-конфигурируемых средств управления отображением информации, допускающих произвольное структурирование статической информации в виде разделов и страниц web-сайта, динамическое формирование разделов сайта, наполнение которых осуществляется из внешних источников, расширение структуры сайта за счет подключения новых источников;

– использование технологии сквозной авторизации (Single Sign On, SSO), единой для всех подсистем Информационного web-портала;

– поддержка протоколов синдикации контента (Rich Site Summary, Atom Syndication Format);

– поддержка открытых интерфейсов для взаимодействия с унаследованными и вновь разрабатываемыми системами.

К основным архитектурным принципам программного решения, описанного в данной главе, относятся:

– применение многозвенной архитектуры;

– разделение визуального оформления и информационного содержания web-страниц;

– использование web-интерфейса как основного инструмента администрирования системы;

– публикация части программных интерфейсов посредством web-сервисов для обеспечения расширяемости системы и возможности использования ее функций сторонними информационных системами;

– проектирование системы, как набора компонентов с четко описанными интерфейсами и поведением, что позволяет разрабатывать ее компоненты независимо, абстрагируясь от особенностей реализации конкретных подсистем;

– использование современных стандартов (SQL, XML, SOAP, WSDL, RSS);

– использование для разработки готовых каркасов приложений (Microsoft .NET);

– полная независимость и самостоятельность решения;

– использование компонентного подхода как при отображении информации, так и при работе с информационными источниками;

– использование единого механизма для работы с разнородными информационными источниками, позволяющего унифицировать доступ и исключить зависимость компонентов CMS от особенностей работы с конкретными источниками;

– наличие методологии и развитого инструментария, позволяющих расширять состав компонентов системы и состав информационных источников.

Далее в главе для поставленных технических требований предложены следующие варианты их практической реализации:

– применение шаблонов для разделения визуального оформления и информационного содержания web-страниц;

– применение в шаблонах визуальных компонентов для отображения встроенных типов контента и типовых элементов пользовательского интерфейса;

– применение виртуализации путей web-страниц для обеспечения неизменности путей при изменении структуры страниц web-сайта;

– использование собственного скриптового языка для настраиваемого отображения в пределах одной страницы данных из различных информационных источников;

– организация работы с информационными источниками через общий сервер интеграции данных, обеспечивающий виртуализацию схемы данных и независимость программной реализации CMS от специфики работы с конкретными источниками;

– применение метаописания, хранящегося в едином конфигурационном файле для управления структурой страниц web-сайта;

– применение механизма модулей расширения для динамического формирования структуры выделенных разделов сайта и управления их представлением;

– организация взаимодействия подсистемы безопасности с выделенным сервисом аутентификации, обеспечивающим независимость программной реализации CMS от специфики используемого механизма аутентификации, с одновременным делегированием принятия решений по вопросам авторизации отдельным подсистемам и информационным источникам.

Перечисленные предложения воплощены в архитектуре программного решения, представленной на рис. 1.

В третьей главе рассмотрены вопросы моделирования среды и процессов функционирования CMS. Основная задача Информационного web-портала – организация доступа пользователей к контенту из собственного хранилища портала, различных информационных источников и служб. Поведение пользователей – это важная составляющая среды функционирования портала. Активность пользователей и состав требующихся им информационных источников не могут быть определены заранее. Также для активности пользователей характерно:

– наличие периодического изменения активности, связанного с тематической направленностью портала. Например, сайты государственных учреждений имеют максимум посещаемости в рабочие дни. В реальных наблюдениях можно видеть суточные изменения активности (день / ночь), недельные (рабочие / выходные дни) и годичные;

– наличие спонтанных всплесков активности, связанных с появлением на ресурсе контента, вызывающего интерес непосредственной

аудитории сайта, либо с появлением ссылки на этот контент на другом высоко посещаемом ресурсе.

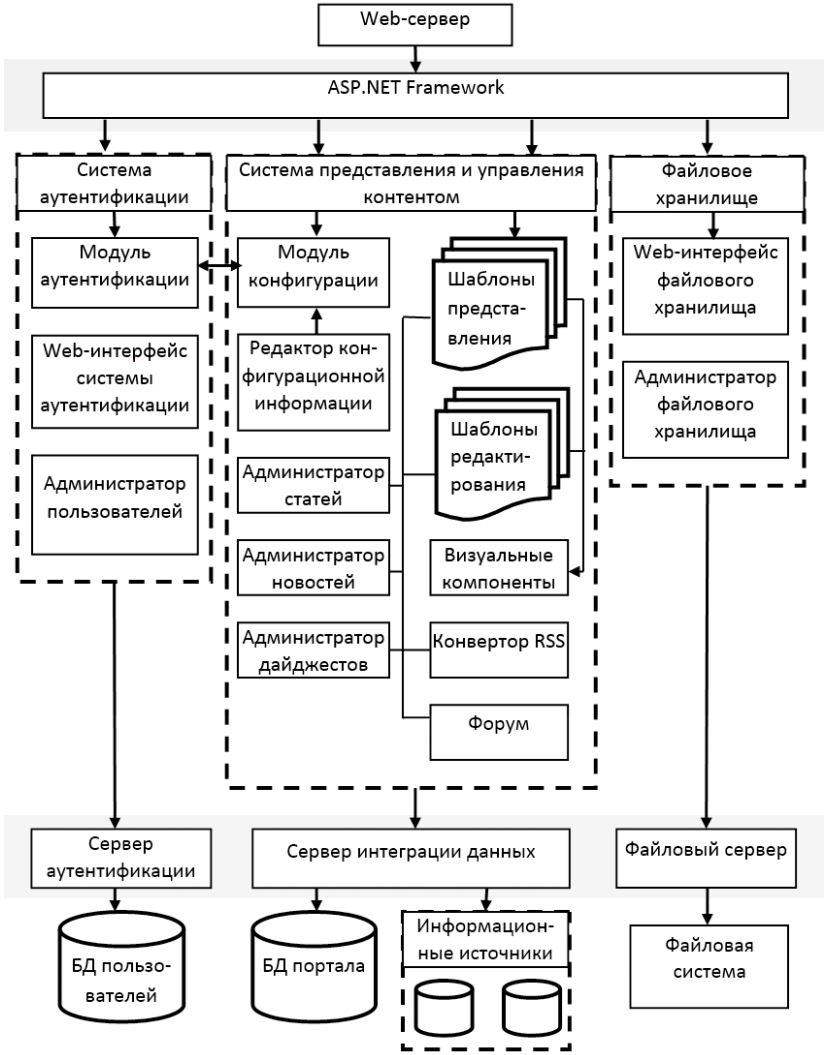

Рис. 1

Значительная часть контента в web-портале размещена в подключенных к нему внешних информационных источниках. При доступе к этому контенту портал, получив исходный web-запрос, трансформиру-

ет его в запросы к источникам, получает от источников контент, как результат выполнения запроса, и передает его пользователю. В связи с этим, источники являются второй значимой составляющей среды функционирования. Таким образом, наиболее важными показателями функционирования портала в целом и CMS в частности являются активность пользователей и эффективность выполнения запросов информационными источниками.

Для исследуемых показателей сложно предложить модели, построенные на основе физических аналогий, так как неизвестны такие физические законы, которые описывали бы интерес пользователей к ресурсу. Вместо этого предлагается сформировать класс моделей, учитывающих характерные (реально наблюдаемые, известные из опыта эксплуатации) поведенческие черты указанных показателей, а затем, оставаясь в рамках предложенного класса моделей, подобрать параметры (идентифицировать модель).

Исследуемые показатели являются динамическими и носят случайный характер, поэтому наиболее простым вариантом для их описания представляется использование линейных стохастических моделей. Однако в реально наблюдаемых данных обнаруживаются зависимость возмущений от текущих значений показателя, цикличность, скачкообразные изменения. Такое поведение с трудом поддается описанию с помощью линейных моделей. В то же время, в наблюдаемых данных преобладают периоды стационарности (когда показатель явно меняется мало и делает это линейным образом), перемежаемые достаточно кратковременными скачкообразными переходными процессами. Таким образом, естественно предполагать наличие нескольких, близких к стационарным, режимов функционирования.

Следует отметить наблюдаемую взаимосвязь между различными режимами функционирования. Например, в режиме малой активности пользователей случается постепенное повышение интереса, которое при достижении некоторой границы скачкообразно сменяется режимом повышенной активности, который, в свою очередь, через некоторое время снова «сваливается» в малоактивный режим.

Для моделирования описанного нелинейного поведения показателей предлагается использовать подход, основанный на квантовании возможных состояний. А именно, предполагается, что пространство значений показателя может быть разбито на области, внутри которых его динамика описывается простейшими линейными уравнениями, а при выходе показателя за границы области модель изменяется. Такой подход, по-видимому, впервые, предложен в работе Tong H., Lim K.S.[4],

[4] Tong H., Lim K.S. Threshold autoregression, limit cycles, and cyclical data. *Journal of the Royal Statistical Society. Series B*, Vol. 42, No. 3(1980), pp. 245-292

где рассматривается модель авторегрессии с порогом (TAR, threshold autoregression).

С помощью модели TAR возможно учитывать смену характера динамики временного ряда (переключение между разными авторегрессионными моделями) за счет введения некоторых пороговых значений. Интервал пороговых значений, в котором оказывается временной ряд в заданные моменты времени, и будем называть режимом. Таким образом, модель исходит из разбиения диапазона значений показателя на конечное число областей, каждая из которых принимается за определенный режим функционирования, и эти режимы могут меняться. Предлагаемый идеологией TAR подход очень важен, т.к. позволяет не привлекать некие «внешние» неконтролируемые силы, меняющие режим, а использовать для этого сам показатель: достижение им одного из пороговых значений приводит к переходу в другой режим функционирования и смене параметров модели.

Непосредственно такой подход довольно легко применить к задаче описания пользовательской активности. Однако возникает сложность использования упомянутой модели TAR применительно к параметру состояния источника. Само понятие состояния источника плохо формализуемо, а значение этого параметра не может быть выяснено непосредственно. При выполнении запросов порталом состояние источника характеризует его типичное на некотором временном интервале время выполнения запроса. Таким образом, судить о состоянии источника возможно только по косвенным признакам, исходя из истории наблюдений за ним. В связи с этим, предлагается моделировать не непосредственно параметр состояния источника, а ожидаемое среднее время выполнения запросов. Наличие переключений параметров модели, используемых в модели TAR, в данном случае помогает формализовать понятие состояния источника: режимы с малым значением времени выполнения запроса источником свидетельствуют о благоприятных условиях его функционирования (низкой загрузке, высокой доступности вычислительных ресурсов), а режимы медленного выполнения запросов свидетельствуют об обратном (высокой загрузке, низкой доступности ресурсов).

В качестве *модели, описывающей динамику показателя эффективности источника,* предлагается следующая стохастическая динамическая система наблюдений общего вида:

$$\begin{cases} r_t = a_t r_{t-1} + s_t + b_t \psi_t, t = 1,2, \dots, \\ m_t = D(r_t) + E(r_t)\varphi_t, \end{cases} \quad (1)$$

где $\{a_t\}, \{s_t\}, \{b_t\}$ – заданные числовые последовательности, $\{\psi_t\}$ – последовательность независимых одинаково распределенных случайных величин, r_0 – случайная величина, не зависящая от $\{\psi_t\}$, $\{\varphi_t\}$ – после-

довательность независимых одинаково распределенных случайных величин (предполагается, что она не зависит от $\{\psi_t\}$ и r_0). Область значений R^1 процесса r_t разбита на непересекающиеся интервалы точками $d{:}-\infty = d_0 < d_1 < \cdots < d_{n-1} < d_n = +\infty$, n – число режимов эффективности, функции $D(x)$ и $E(x)$ определены выражениями

$$D(x) = \begin{cases} D_1, -\infty < x < d_1, \\ D_2, d \le x < d_2, \\ \dots, \\ D_n, d_{n-1} \le x < +\infty, \end{cases}$$
$$E(x) = \begin{cases} E_1, -\infty < x < d_1, \\ E_2, d_1 \le x < d_2, \\ \dots, \\ E_n, d_{n-1} \le x < +\infty \end{cases} \quad (2)$$

Показатель r_t будем интерпретировать как некоторое объективное значение относительной эффективности источника в момент времени *t*. Под этой характеристикой понимается отношение времени затраченного источником на выполнение запросов на текущем интервале наблюдения $(t_{n-1}; t_n]$ к длительности этого интервала. Использование относительной величины призвано исключить влияние разнородности источников и выполняемых ими запросов.

Наличие процесса r_t связано с необходимостью моделирования текущего режима источника. Данный процесс необходим, чтобы обеспечить последовательность смены режимов, адекватную наблюдаемым данным. Смена режима происходит при выходе значений процесса за границы интервала, определенного для выбранного режима. В данном контексте величина и размерность показателя r_t не имеет значения.

Показатель m_t – моделируемое (фактически наблюдаемое) абсолютное время выполнения запросов источником. Функции $D(r_t)$ и $E(r_t)$ – соответственно возвращают ожидаемые средние величины наблюдений и отклонения наблюдаемых величин от средних значений для текущего режима, определяемого значением процесса r_t.

В качестве *модели, описывающей динамику показателя активности пользователей,* предлагается следующая стохастическая динамическая система общего вида:

$$u_t = F(u_{t-1})u_{t-1} + G(u_{t-1}) + B(u_{t-1})\gamma_t, t = 1,2, \dots, \quad (3)$$

где $\{\gamma_t\}$ – стандартный дискретный белый шум, u_0 – случайная величина, не зависящая от $\{\gamma_t\}$, функции $F(x), G(x)$ и $B(x)$:

$$F(x) = \begin{cases} F_1, -\infty < x \le d_1, \\ F_2, d < x \le d_2, \\ \dots, \\ F_n, d_{n-1} < x < +\infty, \end{cases}$$
$$G(x) = \begin{cases} G_1, -\infty < x \le d_1, \\ G_2, d < x \le d_2, \\ \dots, \\ G_n, d_{n-1} < x < +\infty, \end{cases} \quad (4)$$
$$B(x) = \begin{cases} B_1, -\infty < x \le d_1, \\ B_2, d < x \le d_2, \\ \dots, \\ B_n, d_{n-1} < x < +\infty \end{cases}$$

Показатель u_t здесь – общее число выполненных порталом запросов на текущем интервале наблюдения $(t_{n-1}; t_n]$. Следует обратить внимание, что именно число запросов является показателем пользовательской активности, а не число пользователей. Тем самым учитывается, возможность выполнения пользователем нескольких запросов, а также сложность (число задействованных источников) запроса. Данная модель также построена в предположении о наличии режимов интенсивности поступления запросов в портал. В пределах установившегося режима используется обычная авторегрессия. Для смены режимов используется значение самого процесса u_t, а не вспомогательного процесса, как в случае модели показателя эффективности источника (1), поскольку активность пользователей может быть оценена и смоделирована непосредственно.

Функции $F(x)$ и $G(x)$ – определяют ожидаемые средние величины наблюдений, а функция $B(x)$ возвращает отклонения наблюдаемых величин от средних значений для текущего режима, определяемого значением процесса u_t.

Таким образом, для конкретной реализации портала, необходимо выбрать интервалы значений режимов функционирования и подобрать параметры моделей (1) и (3). Значения этих параметров индивидуальны для каждой реализации и существенно зависят от многих факторов: параметров быстродействия серверов, дисковых накопителей, способов реализации информационных источников, целевой аудитории портала. В связи с этим, не меньшую важность, чем выбор модели представляет процедура выбора параметров.

Исследования, проведенные в работе в связи с реализацией этой процедуры, основаны на результатах наблюдения за порталом Российской академии наук www.ras.ru. Для этого ресурса был собран и обобщен существенный объем статистических данных. Также для этого

ресурса известны целевая аудитория, состав и характеристики информационных источников.

Возможны различные подходы к задаче определения параметров моделей (1) и (3). При формальном подходе необходимо поставить задачу идентификации параметров и вычислить оценки, задавшись каким-либо критерием оценивания. Его реализация приведена в работе А.В. Босова[5]. Однако полученные результаты имеют определенные недостатки. Во-первых, были введены дополнительные предположения о параметрах (например, их распределения), обоснованность которых не ясна. Во-вторых, сложность решения таких задач чрезвычайно высока – и с математической, и с вычислительной точек зрения.

В данной работе сделана попытка подобрать параметры моделей, опираясь на простейший статистический анализ. На основе реального набора данных производится выбор режимов. Затем из исходных данных формируются выборки, соответствующие нахождению исследуемого показателя в выбранном режиме. Для выборок оцениваются основные статистические параметры (среднее, дисперсия). В предположении, что в рамках установившегося режима среднее значение и дисперсия процесса остаются неизменными, становится возможным вычислить параметры модели через простейшие уравнения. После этого выполняется моделирование и оценка адекватности модели исходным данным.

Таким образом, устанавливается возможность обоснованного выбора параметров рассматриваемых моделей за счет простейшего статистического анализа среды функционирования портала, по крайней мере, в случае портала РАН. Собственно, действия, выполняемые при этом выборе можно рассматривать в качестве методики, применимой и к другим реализациям портала. Данный подход назван экспертным оцениванием, т.к. задача разделения данных между режимами выполняется экспертом.

В работе описаны примеры применения процедур экспертного оценивания параметров на основе экспериментальных данных, собранных в процессе функционирования портала http://www.ras.ru/.

На рис. 2 приведены данные среднего времени выполнения запросов одним из источников портала – файловым хранилищем (FStorage). Выделены 3 режима функционирования источника: с высокой, средней и низкой эффективностью. К режиму высокой эффективности были отнесены наблюдения со значением менее 0,2 секунды, к режиму низкой эффективности наблюдения со значением более 0,6 секунды, к режиму средней эффективности – наблюдения в интервале от 0,2 до 0,6

[5] Босов А.В. Моделирование и оптимизация процессов функционирования Информационного web-портала // Программирование, № 6, 2009.С.53-66.

секунды. Затем интервалы были сделаны непересекающимися, чтобы исключить значения, находящиеся вблизи границ, и которые трудно однозначно отнести к тому или иному классу. Уточненные интервалы составили (0; 0,15], [0,2; 0,45] и [0,6; +∞).

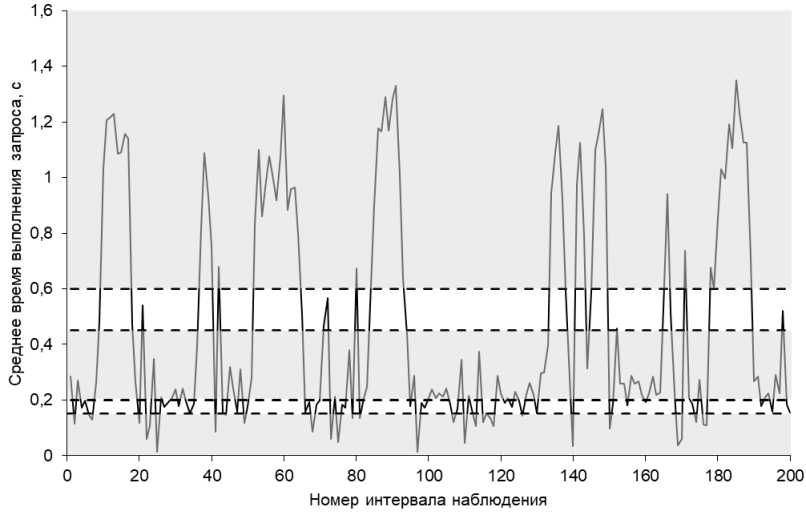

Рис. 2

Теперь остается выбрать такие параметры a_t, s_t, b_t, чтобы распределение значений процесса соответствовало фактическим частотам нахождения источника в выбранных режимах. Подбирались эти параметры постоянными, так как оснований для выделения каких-то «особых» моментов времени t нет. Параметр $a = a_t$ характеризует влияние предыстории на его текущее и будущее состояния. Если для источника характерно долговременное пребывание в рассматриваемом режиме, параметр a выбирается близким к единице.

Параметры $s = s_t$ и $b = b_t$ выбираются в предположении о наличии стационарного режима функционирования портала. Данное предположение подразумевает, что среднее значение процесса r_t не должно изменяться, т.е. $\boldsymbol{M}[r_{t_k}] = \boldsymbol{M}[r_{t_{k-1}}] = s/(1-a)$. Также должна оставаться постоянной и дисперсия r_{t_n}, т.е. $\boldsymbol{D}[r_{t_k}] = \boldsymbol{D}[r_{t_{k-1}}] = b^2/(1-a^2)$.

Полученные модели могут быть использованы при решении целого ряда практических вопросов, возникающих при разработке и внедрении таких web-систем, как информационный web-портал: настройке параметров производительности, анализа работоспособности, оценке эффективности функционирования под нагрузкой и т.п. Традиционно, решение таких вопросов приходится выполнять в условиях реально

функционирующей системы, поскольку в условиях стенда может быть достоверно смоделирована только конфигурация программных и аппаратных средств. В условиях стенда не составляет трудности смоделировать пользовательские запросы, однако характеристики потока запросов выбираются обычно произвольно и соотносятся с реальной нагрузкой весьма условно. Произвольные значения имеют и характеристики информационных источников при отсутствии реального информационного наполнения. При этом работа под реальной нагрузкой существенно усложняет как исследование работы web-системы, так и внесение изменений в ее работу, поскольку сопровождается снижением доступности системы для конечных пользователей. Применение же предложенных моделей позволяет в условиях стенда получить характеристики процессов пользовательской активности и функционирования информационных источников, близкие к наблюдаемым при работе системы с реальной нагрузкой. За счет изменения параметров соответствующих моделей становится возможным моделировать различные виды информационных источников и режимы пользовательской активности. Как результат, появляется возможность анализировать работоспособность и эффективность web-системы с различными типами информационных источников и режимами пользовательской активности, при различных комбинациях настроек.

Аналогичная методика применяется и для построения модели пользовательской активности (3). В диссертационной работе приведены исчерпывающие примеры.

Четвертая глава посвящена примерам практического применения разработанной CMS, а также применению полученных в главе 3 моделей к задаче оценивания вероятностно-временных характеристик программного обеспечения (ПО).

В перечень проектов входят:

1) Информационный Web-портал Российской академии наук, созданный в рамках программы Президиума РАН «Информатизация» (2001-2010 гг.). Информационный Web-портал является официальным представительством Российской академии наук в сети Интернет. Web-портал используется как основное средство интеграции наследуемых информационных ресурсов, в число которых входят общеакадемические информационные хранилища по основным категориям организационно-административных данных. Web-портал РАН также рассматривается как основное средство интеграции основных взаимосвязанных категорий научных цифровых ресурсов РАН, взаимодействия с имеющимися региональными и зарубежными информационными системами, с информационными системами РФФИ, Министерства образования и науки, органов государственной власти и управления, российских ВУЗов.

2) Учебный портал, созданный в 2008-2010 гг. на базе имеющихся решений по управлению контентом. Учебный портал – это система дистанционного обучения, предназначенная для организации учебного процесса в форме дистанционного обучения через сеть Интернет или Интранет с использованием современных образовательных технологий. Система образована подсистемами управления обучением и управления учебным контентом. Подсистема управления обучением – это средство организации учебного процесса, контроля знаний и решения различных административных задач, включая регистрацию обучаемых, назначения им конкретных курсов, сбора статистики обучения. Подсистема управления учебным контентом является средством для создания учебного контента, т.е. учебных и информационных материалов из которых формируются учебные курсы.

Задача оценивания вероятностно-временных характеристик ПО предполагает сбор и последующий статистический анализ данных о работе ПО, накапливаемых пользователями на этапе опытной эксплуатации. Ограниченность во времени этого этапа, отсутствие у пользователей навыков тестирования ПО, сложность создания прецедентов пиковой нагрузки затрудняют получение необходимого объема данных.

Предлагаемое решение состоит в использовании имитационного моделирования пользовательской активности. Для этого была разработана программа-имитатор, позволяющая имитировать взаимодействие пользователя с тестируемым ПО и реализующая модель пользовательской активности, описанную в главе 3. Это позволило оценивать надежность ПО без привлечения к деятельному участию в этом пользователей. При этом были достигнуты и другие цели. Например, данная методика хорошо зарекомендовала себя как способ нагрузочного тестирования программ.

В **заключении** приведены основные итоги диссертационной работы и сформулированы результаты, представляемые диссертантом к защите.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

Основными научными и практическими результатами диссертационной работы являются:

– проектные решения по архитектуре и программной инфраструктуре портальной CMS, полученные с использованием методов объектно-ориентированного проектирования и анализа алгоритмов и программ;

– программа для ЭВМ, обеспечивающая управление контентом и предназначенная для использования в составе интеграционной платформы информационного web-портала.

– математическая модель динамики показателя пользовательской активности, представленная стохастической динамической системой, использующей авторегрессионную модель с переключениями;

– математическая модель динамики показателя эффективности информационного источника, представленная стохастической динамической системой наблюдений, использующей авторегрессионную модель с переключениями. Предложен подход к определению параметров модели на основе простейшего статистического анализа среды функционирования web-портала;

– алгоритмические и программные решения для нагрузочного тестирования и анализа вероятностно-временных характеристик программной системы на основе интеграционной платформы Информационного web-портала.

ОСНОВНЫЕ ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Соколов И.А., Босов А.В., Зацман И.М., Иванов А.В., Чавтараев Р.Б. О концептуальных основах разработки Единой информационной системы РАН // Системы и средства информатики. Вып. 12. – М.: Наука, 2002. С. 29-47.

2. Босов А.В., Иванов А.В. О реализации системы управления содержанием информационного Web-портала // Информационные технологии и вычислительные системы. №4. – М.: ИМВС РАН, 2004. С.85-103.

3. Босов А.В., Иванов А.В. Технология управления содержанием в информационном портале РАН // Системы и средства информатики. Вып. 15. – М.: Наука, 2005. С.260-283.

4. Босов А.В, Иванов А.В., Полухин А.Н., Чавтараев Р.Б. Управление сайтом информационного web-портала // Системы и средства информатики. Вып. 15. – М.: Наука, 2005. С.233-259.

5. Босов А.В., Иванов А.В. Программная инфраструктура Информационного web-портала РАН // Информатика и ее применения, Вып.2, Т.1, 2007, С.39-53.

6. Иванов А.В. Математические модели базовых процессов функционирования Информационного web-портала // Системы и средства информатики. Выпуск 20, №1. / Под ред. И.А. Соколова. – М.: ТОРУС ПРЕСС, 2010. С.106-132.

7. Борисов А. В., Босов А. В., Иванов А.В., Корепанов Э. Р. К вопросу расчета надежности информационно-телекоммуникационных

систем: учет характеристик программного обеспечения // Системы и средства информатики, 2018. Т. 28. № 1. С.20-34.

8. Борисов А. В., Босов А. В., Иванов А. В., Чавтараев Р. Б. Имитационное моделирование пользовательской активности для оценивания вероятностно-временных характеристик программного обеспечения // Системы и средства информатики, 2018. Т. 28. № 2. С.20-34.

9. Иванов А.В., Босов А.В., Чавтараев Р.Б. Инструменты интеграции Информационного web-портала РАН // Тезисы докладов. XII Российская конференция с международным участием «Распределенные информационно-вычислительные ресурсы» (DICR2008), Академгородок, Новосибирск, 5-7 ноября 2008, С.21-22.