На правах рукописи

НОВИКОВА ОЛЬГА АЛЕКСАНДРОВНА

РАЗРАБОТКА И ИССЛЕДОВАНИЕ МОДЕЛЕЙ ПРОГНОЗИРОВАНИЯ ДИНАМИКИ НОВОСТНЫХ ЛЕНТ

Специальность 05.13.17 – «Теоретические основы информатики»
(по техническим наукам)

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Москва-2020

Диссертация выполнена на кафедре «Корпоративные информационные системы» (КИС) Института информационных технологий (ИИТ) Федерального государственного бюджетного образовательного учреждения высшего образования «МИРЭА - Российский технологический университет» (РТУ МИРЭА)

Научный руководитель:

Жуков Дмитрий Олегович

доктор технических наук, профессор кафедры КБ-8 Института комплексной безопасности и специального приборостроения Федерального государственного бюджетного образовательного учреждения высшего образования «МИРЭА – Российский технологический университет» (РТУ МИРЭА)

Официальные оппоненты:

Барахнин Владимир Борисович

доктор технических наук, доцент, ведущий научный сотрудник лаборатории цифровых двойников и анализа больших данных Федерального государственного бюджетного научного учреждения «Федеральный исследовательский центр информационных и вычислительных технологий» (ФИЦ ИВТ)

Загорулько Юрий Алексеевич

кандидат технических наук, заведующий лабораторией искусственного интеллекта Института систем информатики им. Ершова Сибирского отделения Российской академии наук (ИСИ СО РАН)

Ведущая организация:

Федеральное государственное автономное образовательное учреждение высшего образования «Уральский федеральный университет имени первого Президента России Б.Н. Ельцина» (УрФУ)

Защита состоится «04» марта 2021 года в 14 часов 00 мин. на заседании диссертационного совета Д 219.005.02 в Федеральном государственном бюджетном образовательном учреждении высшего образования «Сибирский государственный университет телекоммуникаций и информатики» (СибГУТИ) по адресу: 630102, г. Новосибирск, ул. Кирова, 86 ауд. 625.

С диссертацией можно ознакомиться в библиотеке СибГУТИ и на сайте: http://www.sibsutis.ru/science/postgraduate/dis_sovets/.

Автореферат разослан «___» _____ 20 г.

Ученый секретарь
диссертационного совета,
к.т.н., доцент **И.В. Нечта**

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность исследования. В последние годы сведения, которыми мировые информационные агентства заполняли свои новостные электронные ленты, стали доступны широкому потребителю. Более того, информация практически в любой области человеческой деятельности вышла за рамки обладания и работы с нею лишь узким кругом специалистов. Теперь каждый индивидуум, владеющий техническими возможностями выхода в мировые информационные сети, может вносить собственную лепту в формирование глобальной копилки знаний. При таких масштабах большая часть этих данных содержится в хаотичном или слабо организованном виде, что, в свою очередь, привело к необходимости совершенствования средств их структурирования и анализа, а также поиска методов извлечения закономерностей в корпусе новостных сообщений и прогнозирования вероятности появления в новостной ленте сообщений о значимых событиях с целью упреждающего воздействия на различные виды деятельности и управления их возможными состояниями. Например, зная о том, что с большой вероятностью через определенный промежуток времени могут произойти гражданские беспорядки, можно предпринять ряд мер для того, чтобы не допустить их появления или избежать серьёзных последствий.

Разработка и совершенствование средств анализа и выявления закономерностей в текстовых данных, в том числе и в новостных сообщениях, исследование особенностей динамики новостных лент являются важными и актуальными научными задачами, требующими для своего решения поиска новых подходов, основанных на информационных технологиях обработки больших данных. Существенный вклад в разработку решений данной проблематики внесли: Хадйэ Нассиртоусси (Khadjeh Nassirtoussi), Л. Чжао (L.Zhao), Н. Рамакришнан (N.Ramakrishnan), Г. Финние (G. Finnie), Д.О. Жуков, Л. Анастасакис (L. Anastasakis), Ц.-Й. Хуанг (C.-J. Huang), Б. Ванстоне (B. Vanstone), Д. А. Замотайлова, Е. Лупиани-Руиз (E. Lupiani-Ruiz), И. Гарциа-Манотас (I. García-Manotas), Й. Клеиннийенхуис (J. Kleinnijenhuis), Х. Ю (H. Yu), М.Либман (M. Liebmann), К. Яковлева и многие другие.

Отличительной особенностью новостного пространства, в силу наличия человеческого фактора, является стохастический характер протекающих в нем процессов, возможность самоорганизации информации и наличие памяти. Слабоструктурированность информационного пространства новостных текстов является одной из основных проблем в прогнозировании динамики новостных лент. Поэтому актуально создать модель для прогнозирования появления сообщений в новостной ленте, которая учитывала бы наличие человеческого фактора и была бы основана на стохастической динамике изменений структуры новостных кластеров (или состояний информационного пространства), и которая учитывала бы память и самоорганизацию их структуры.

В основе, предлагаемой в данной работе, методики анализа динамики новостных текстов и модели прогнозирования появления события в сообщениях новостной ленты, лежат следующие шаги.

Собирается корпус новостных текстов за большой промежуток времени (например, за 15-20 лет). Так как новостные сообщения в сети Интернет представляют собой гетерогенные данные (данные, имеющие различную форму представления и неодинаковые единицы измерения), они могут включать в себя текст, числовые данные, денежные знаки, время/дату, гиперссылки и др. В одной математической модели невозможно одновременно производить вычисления над различными типами данных, а значит, необходим такой математический аппарат, который позволил бы привести гетерогенные данные, составляющие основу новостных сообщений, к единой шкале измерений. Чтобы это осуществить, необходимо выполнить соответствующие вычислительные операции по отображению данных на числовое множество:

- выполнить лингвистическую обработку тестов новостных сообщений (удаление знаков препинания, удаление стоп слов, нормализацию, лемматизацию);
- создать словарь, на основе которого осуществляется векторизация всех новостных сообщений. В связи с тем, что каждое новостное сообщение может быть описано конечным набором терминов, из этих наборов терминов можно создать новостные векторы в

информационном пространстве. Координаты новостного вектора будут представлять собой количество вхождений терминов из словаря в текст новостного сообщения;

- кластеризовать полученные векторы новостного пространства по смысловым группам, осуществить уточнение кластеризации.

Далее формируются временные ряды изменения структуры полученных кластеров с течением времени (например, в качестве параметра при создании временного ряда, можно взять изменение положения центров кластеров с течением времени. Как и любой временной ряд, данные ряды могут содержать циклические колебания, тренд, сезонные колебания и стохастическую компоненту.

Достаточно хорошо исследованы методы анализа временных рядов, в которых сохраняется тренд или имеющих сезонную составляющую (метод Хольт-Винтерса, R/S-анализ, алгоритм ARIMA и др.). Когда временной ряд в большей степени обладает стохастической компонентой, задача становится намного сложнее. Если значение переменной Хёрста (например, определенной в результате R/S-анализа) находится вблизи 0,5, наблюдается некоррелированное поведение ряда, процесс является стохастическим. Одним из возможных решений в таком случае может стать подход, построенный на гипотезе о том, что существуют причинно-следственные связи между событиями, происходящими в реальном мире. А, это значит, что можно построить математическую модель, отображающую связь между новостными векторами уже произошедших событий (событий, описанных в сообщениях новостной ленты) и вектором прогнозируемого новостного сообщения (вектором, созданным с помощью словаря, используя текстовое описание прогнозируемого события).

Суть такой модели может заключаться в следующем: сначала создаем текстовое описание образа новостного события, для которого необходимо определить вероятность его реализации с течением времени (прогноз). Далее векторизуем текстовое описание прогнозируемого события (получаем вектор X_{bs}). Затем определяем, для какого-либо момента времени t значения косинусов углов между векторами центроидов и вектором прогнозируемого события, вычисляем их среднее значение. Величина среднего значения косинусов в данный момент времени будет являться точкой на числовом отрезке [0,1], и, вследствие изменения структуры кластеров с течением времени, она будет совершать на нем почти случайные перемещения (блуждания). С течением времени она может достигнуть заданного значения косинуса, которое мы будем считать порогом реализации события (назовем его l). Текущую величину среднего значения косинусов назовем состоянием информационной системы в данный момент времени (обозначим его x_0). Вероятность достижения порога события l будет зависеть от времени t (то есть, по сути, мы рассматриваем почти случайные блуждания точки на отрезке [0,1], который содержит в l ловушку, куда с течением времени может попасть блуждающая точка).

Изложенный подход позволяет, на основе рассмотрения схем вероятностных переходов между состояниями информационной системы, сформулировать краевую задачу о зависимости вероятности достижения прогнозируемого сообщения новостной ленты от времени и рассмотреть её решение, на основе модели учитывающей память о предыдущих состояниях информационной системы и их возможную самоорганизацию.

Целью диссертационного исследования является разработка математической модели прогнозирования динамики новостных лент на основе динамики формирования вектора новостного сообщения из векторов информационного новостного пространства текстовых документов, отличительной особенностью которого является стохастический характер протекающих в нем процессов, наличие памяти о предыдущих состояниях и возможность самоорганизации информации.

В качестве технологического инструмента работы с новостным текстом будут применены методы и алгоритмы семантического анализа текстовой информации, позволяющие работать с векторами в качестве математических объектов и формировать временные ряды изменения структуры кластеров.

Для достижения поставленной цели должны быть решены следующие **основные задачи**:

1. Провести анализ современных исследований и разработок в области выявления

закономерностей в текстовых данных и прогнозирования контента новостных лент.

2. Разработать стохастическую модель прогнозирования динамики новостной ленты:
   - построить разностные схемы для вероятностей переходов между состояниями информационной системы, описывающими эволюцию рассматриваемого процесса с течением времени. При описании процессов перехода между состояниями учесть возможность самоорганизации и наличие памяти;
   - используя аппарат теории вероятностей и метод графических диаграмм переходов между состояниями, получить алгебраическое уравнение, описывающее условные вероятности соответствующих переходов между возможными состояниями информационной системы;
   - используя аппарат классического математического анализа, с помощью разложения в ряд Тейлора членов вероятностного алгебраического уравнения, получить дифференциальное уравнение второго порядка, описывающее поведение системы (функции зависимости плотности вероятности от времени);
   - сформулировать граничные и начальные условия для краевой задачи, решение которой будет описывать процесс перехода между состояниями в информационном пространстве, решить краевую задачу (например, с помощью преобразований Лапласа) и проанализировать полученное решение.
3. Разработать методику применения разработанной модели для прогнозирования появления события в сообщении новостной ленты на основе его формирования из существующих событий новостной ленты.
4. Провести экспериментальную проверку модели прогнозирования динамики новостных лент и разработать методику оценки адекватности прогностической модели по экспериментальным результатам.

Объектом исследования является поток событий новостной ленты.

Предмет исследования определяется паспортом специальности 05.13.17, область исследования №5 (разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечения, разработка и исследование методов и алгоритмов анализа текста, устной речи и изображения).

Методология и методы диссертационного исследования. Научные положения диссертации получены с использованием моделирования систем, теории систем и системного анализа, методов математического анализа, теории вероятностей и дифференциального исчисления, операционного исчисления, методов математической лингвистики, теорий классификации и систематизации, теории самоподобия.

Научная новизна. Научная новизна обусловлена тем, что в представленном диссертационном исследовании:

1. Разработан подход к моделированию процессов прогнозирования появления событий новостной ленты, основанный на возможности использования схем вероятностей переходов между состояниями в информационном новостном пространстве с учетом памяти о предыдущих состояниях; выводе алгебраического и дифференциального уравнений для описания условных вероятностей переходов с течением времени; формулировке и решении граничной задачи.
2. Выведено аналитическое выражение для нахождения зависимости от времени плотности вероятности обнаружения системы в одном из возможных состояний.
3. Разработана методика прогнозирования появления события новостной ленты на основе анализа изменения структуры кластеров с течением времени и вероятностной модели формирования событий в новостной ленте с учетом памяти о предыдущих состояниях информационной системы и самоорганизации информации.

Теоретическая и практическая значимость.

Теоретическая значимость состоит в том, что разработана новая математическая модель, в которой рассматриваются возможные переходы между состояниями информационной системы с учетом предыдущих шагов (или состояний), на основе чего было выведено алгебраическое уравнение, описывающее условные вероятности перехода между состояниями, разложение

которого в ряд Тейлора позволяет получить дифференциальное уравнение, учитывающее не только первые, но и вторые производные, что позволяет говорить о самоорганизации информационной системы, а на основе данного уравнения была сформулирована и решена краевая задача.

Практическая значимость диссертационного исследования заключается в том, что на основе разработанной теоретической модели можно создать алгоритм, с помощью которого возможно прогнозирование появления событий в тех случаях, когда не работают классические методы анализа и прогнозирования временных рядов, такие как, например, R/S-анализ. Результаты экспериментов показали, что разработанная модель является адекватной и не противоречивой.

Результаты работы использовались в конкурсной части государственных заданий высшим учебным заведениям и научным организациям по выполнению инициативных научных проектов, финансируемых Министерством образования и науки РФ (проект № 28.2635.2017/ПЧ «Разработка моделей стохастической самоорганизации слабоструктурированной информации и реализации памяти при прогнозировании новостных событий на основе массивов естественно-языковых текстов».

Обоснованность и достоверность научных положений, основных выводов и результатов диссертации обеспечивается за счет положительной экспериментальной проверки, корректного использования математического аппарата, который был применён для получения основных уравнений модели. Полученные результаты и выводы непротиворечивы результатам и выводам, которые можно сделать на основе анализа состояния данной предметной области, а также результатам, полученным в ходе экспериментов, подтверждающими адекватность разработанной модели прогнозирования динамики новостных лент. Основные теоретические положения диссертации апробированы в печатных трудах и докладах на научных конференциях, где результаты исследования не вызвали серьёзных нареканий со стороны научного сообщества.

Внедрение результатов исследования осуществлено в учебную работу на кафедре информационных технологий в государственном управлении (ИТГУ) Института инновационных технологий и государственного управления (ИНТГУ) ФГБОУ ВО «МИРЭА – Российский технологический университет», в практику деятельности компаний ООО «РУСНЕФТЕТРЕЙД» (респ. Башкортостан) и ООО НАУЧНО-ТЕХНИЧЕСКИЙ ЦЕНТР «ЭССЗ» (г. Санкт-Петербург), что подтверждается соответствующими актами.

Апробация результатов работы. Основные теоретические и практические результаты диссертационного исследования апробированы в научно-исследовательских работах и отражены в докладах на научно-практических и научных конференциях: Международная научная конференция Big Data & AI Conference 2020 (Москва, 17-18 сентября 2020 г.), VIII Международная научная конференция "Компьютерные науки и информационные технологии», памяти А.М. Богомолова (Саратов, 2018 г.), конференции «ITM Web of Conferences» (2017 г.), I Международной научной конференции «Конвергентные когнитивно-информационные технологии» (г. Москва, 25–26 ноября 2016 г.), Международная заочная научно-практическая конференция (Тамбов, 29 февраля 2012 г.); в результатах конкурсов молодых ученых: Всероссийский конкурс научно-исследовательских работ студентов и аспирантов в области информатики и информационных технологий (Белгород, 2012 г.), IV Всероссийский конкурс молодых ученых (Москва, 2012 г.).

Кроме того, результаты работы докладывались на научно – технических семинарах ФГБОУ ВО «МИРЭА - Российский технологический университет».

Основные положения, выносимые на защиту:

1. Предложенные научно-обоснованные принципы формирования корпуса новостных текстов и соответствующий набор количественных показателей обеспечивают адекватное отображение контента новостных текстов в цифровое пространство.
2. Динамику изменения контента новостных лент можно описывать на основе результатов анализа временных рядов, составленных из текущих значений количественных показателей изменения параметров корпусов новостных текстов, которые с математической точки зрения представляют собой реализации случайных процессов в

информационном пространстве с учетом самоорганизации и наличия памяти.

3. На основе использования решений предложенного в диссертации вероятностного уравнения, описывающего случайные процессы с самоорганизацией и памятью, можно с точностью не менее 60% (зависит от глубины памяти) вычислять прогнозируемые значения временных рядов, составленных из текущих значений наблюдаемых количественных показателей изменения параметров корпусов новостных текстов.

Личный вклад автора. Представленные в данной диссертационной работе исследования являются результатами работы, проведенной автором диссертации. Основные результаты исследования отражены в 13 научных работах, из них 5 являются публикациями в рецензируемых журналах, рекомендованных ВАК Минобрнауки РФ для опубликования основных результатов диссертационных исследований на соискание ученых степеней доктора и кандидата наук, 3 опубликованы в трудах международных конференций, входящих в базы SCOPUS и Web of Science.

Получено свидетельство о государственной регистрации программы для ЭВМ №2018615544 от 10.05.2018 «Модуль прогнозирования новостных событий на основе анализа спектров информационных процессов».

Диссертация состоит из введения, четырёх глав, заключения, списка использованной литературы, пяти приложений. Общий объём работы с приложениями, 18 рисунками, 5 таблицами – 172 стр.

СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Во введении раскрывается актуальность темы диссертационного исследования, обосновывается выбор предмета и объекта, формулируется цель, определяются основные задачи, которые должны быть решены для достижения поставленной цели исследования, представлены практическая значимость и научная новизна полученных результатов, дана общая характеристика основных положений, выносимых на защиту.

В первой главе анализируются существующие модели и алгоритмы выявления закономерностей в текстовых данных, взятых из новостей, блогов, форумов и т.д. и прогнозирования динамики текстов/новостей с использованием различных инструментов анализа текстовых данных, которые могли бы применяться в решении, сформулированных во введении, задач.

В частности, выполнен анализ работ по прогнозированию изменений рынка ценных бумаг с использованием методов интеллектуального анализа текста. Рассмотрена роль методов интеллектуального анализа текста в автоматизации исследований на основе традиционных методологий прогнозирования: техническом и фундаментальном анализе. Сделан вывод об актуальности и необходимости расширения традиционных методов средствами интеллектуального анализа текстовых и новостных данных, повышающих эффективность прогноза котировок рынка ценных бумаг. Обоснована необходимость учёта данных, извлекаемых из новостных и текстовых сообщений, характеризующих эмоциональную окраску событий, влияющих на изменения рынка (поведенческая экономика, эмоциональный анализ). Сделан обзор применения методов анализа и прогнозирования изменений рынка ценных бумаг, с использованием таких моделей, как машинное обучение, нейронные сети, нечеткая логика, метод опорных векторов регрессии, генетическое программирование сетевых приложений и др.

Также в первой главе проанализированы исследования по прогнозированию в реальном секторе экономики и в социальных системах (рассмотрены исследования по выявлению закономерностей социально-значимых событий). В частности, анализ, извлеченной с помощью технологий обработки больших данных из новостей, информации о чрезвычайных происшествиях, катастрофах и стихийных бедствиях за продолжительный период времени, который может позволить определить интервал времени их самоподобия, что даст возможность прогнозировать частоту их реализации в будущем.

На основании проведенного аналитического обзора современных исследований моделей, алгоритмов анализа данных и разработок в области прогнозирования с использованием текстовых (в том числе новостных) данных, сделан вывод: авторы работ используют алгоритмы анализа и

выявления закономерностей в текстовых/новостных данных и методы прогнозирования появления событий, основанные на большом разнообразии математических моделей, даже если и принимают во внимание стохастический характер протекающих процессов, не учитывают возможности самоорганизации и наличия памяти, что характерно для информационного пространства, представляющего собой корпус новостных текстов. Это может сказаться как на формировании новостной ленты, так и на её динамике.

Сделан вывод, что одним из возможных перспективных направлений в разработке моделей прогнозирования динамики новостных лент может быть создание вероятностной модели формирования образа прогнозируемого события новостной ленты из уже реализовавшихся событий. Такие модели могут быть созданы на основе формализованного математического представления новостных текстов в векторном виде и использования аппарата теории вероятностей для расчета вероятностей возможных состояний информационных процессов в системах управления информационными ресурсами.

В выводах к первой главе предложен ряд основных и дополнительных задач, решение которых позволит достичь цели, поставленной в диссертационном исследовании.

Вторая глава посвящена вопросам обоснования принципов формирования текстового набора данных и выбору количественных показателей для описания динамики новостных текстов.

Рассмотрены вопросы взаимосвязи объектов, явлений, процессов реального мира, причинно-следственные связи, случайность. Сделан вывод о том, что чем выше продолжительность анализируемого интервала, тем больше роль причинной зависимости асимптотически стремиться к нулю, а роль случайности – к единице. А новости, будучи текстовым отражением событий окружающего мира, могут отражать и причинно-следственные связи между событиями. Значит, можно выдвинуть гипотезу о том, что на основании анализа одних новостных текстов можно попробовать прогнозировать появление прогнозируемого события в сообщении новостной ленты. Для этого необходимо разработать модель прогнозирования динамики новостных лент.

Проведён анализ особенностей новостных текстов, отмечена высокая степень их динамичности, перечислены основные характеристики и факторы, относящие новостные тексты к базовыми в СМИ, а также степень влияния на различные сферы жизни: политическую, экономическую, социальную, финансовую, медицинскую и др. Новости способны расколоть или объединить общество, оказать влияние на рынки, могут решать судьбу предприятий, стать причиной протестов и митингов. Сделан вывод о том, что новости не только отражают события, произошедшие в мире, но и сами могут стать причиной новых событий, так как интерпретация получателями смысла новостного сообщения может вызвать ответную реакцию и стать причиной различных действий (спровоцировать митинг, общественный беспорядок), что в свою очередь, будет отражено в новых новостных сообщениях, описывающих эти события.

Также в главе описано обоснование выбора формирования текстового датасета по заданным условиям, которым должны удовлетворять информационные ресурсы в сети Интернет. Среди них: открытость источника информации, периодичность выхода информации с указанием времени выпуска новостной статьи, охват архива материалов, широта аудитории, для которой выпускается информация.

Для прогнозирования событий новостной ленты и построения математических моделей динамики их изменения требуется процедура формализации атрибутов текстов, с целью возможности их измерения в шкалах, применение которых позволяет проводить не только качественные, но и количественные измерения. Так как в одной математической модели невозможно производить вычисления над данными, имеющими разное измерение и формат (измеренными в разных единицах и шкалах), то необходимо выбрать такой математический аппарат, который позволяет привести текстовую информацию, имеющую разное смысловое значение, составляющую основу новостных сообщений, к единой шкале измерений. Сделан обоснованный выбор наиболее подходящей шкалы измерений, соответствующей целям дальнейших исследований, позволяющей выполнить соответствующие вычислительные операции по отображению новостных данных на числовое множество. Это возможно осуществить, если

представить тексты в виде векторов, элементами которых являются числа (т.к. все вектора относятся к одному пространству, что позволяет проводить над ними любые математические операции, которые допускает метрическая или интервальная шкала).

Рассмотрены вопросы предобработки текстовых данных: векторизация, кластеризация новостных сообщений, качества кластеризации, которое в значительной степени определяет дальнейшие результаты обработки текстовой информации, что в конечном итоге сказывается и на качестве анализа динамики новостных данных, так как плохо подготовленные исходные данные не дадут адекватный результат работы математической модели. Предложены методы векторизации, кластеризации и метод повышения точности кластеризации неструктурированных наборов категорийных данных с использованием семантико-энтропийного регулирования, которые позволяют сформировать временные ряды, описывающие события определенного типа в новостных лентах. В результате кластеризации текстовых документов по смысловым группам, из матрицы термин-документ можно выбрать подмножества векторов, каждое из которых образует свой кластер. Можно построить временной ряд с измеряемыми в метрической шкале числовыми характеристиками, взяв за параметр, к примеру, положения центроидов кластеров и рассмотрев их изменения с течением времени, или частоты появления сообщений в новостной ленте, описывающие заданные события. За счет того, что новостные события могут с течением времени появляться и исчезать, структура новостных кластеров и положение векторов, задающих их центры (центроиды) будет изменяться. Полученные кластеры могут быть сегментированы на подгруппы новостей за сутки (24 часа) без суммирования за предыдущие периоды. А затем для каждой из подгрупп может быть определена косинусная метрика между текущим центроидом каждого кластера и базовым вектором (все элементы базового вектора являются постоянными равными единице величинами).

Анализ характеристик динамики полученных временных рядов можно использовать для прогнозирования их эволюции и определения вероятности реализации событий в течении заданного интервала времени, а также отсутствия или наличия в его поведении долговременных зависимостей.

Во второй главе рассмотрена методика формирования временного ряда из контента новостных текстов и постановка задачи моделирования динамики временных рядов изменения контента в новостных лентах.

Полученные результаты показали, что исследуемые временные ряды в общем случае не являются стационарными, а анализ квазистационарных участков наблюдаемых временных рядов и построение выборочных функций распределения чаще всего оказываются малоэффективными для прогнозирования последующей эволюции.

Выделена основная проблема анализа и моделирования поведения временного ряда событий новостной ленты с точки зрения получения прогнозов: в любой момент времени есть только одна реализация процесса (один образец уже реализовавшегося временного ряда), с помощью которой необходимо создать прогноз для следующих моментов времени.

Исследования изменения структуры новостных кластеров с течением времени показали, что существующие приемы изучения динамики изменения временных рядов (например, R/S-анализ), позволяющие в некоторых случаях сделать выводы о наличии или отсутствии долговременных зависимостей, наличии или отсутствии памяти в рассматриваемых процессах, не позволяют спрогнозировать появление конкретного события.

Возникновение событий в новостной ленте связано с наличием человеческого фактора, что вносит неопределенность воздействия на процессы и создает стохастичность, а также создает возможности для их самоорганизации и определяет наличие памяти о предыдущих действиях.

Сделаны выводы, что использование существующих методов анализа временных рядов для моделирования динамики событий новостной ленты, может привести к существенным ошибкам из-за большой изменчивости их характеристик, нелинейности, нестационарности, самоорганизации происходящих процессов и наличия памяти. Потому необходим поиск новых методов анализа их динамики и аппроксимирующих функций распределения. Одним из перспективных направлений создания моделей прогнозирования новостных событий на основе анализа текстовой информации является использование теоретико-вероятностных подходов,

основанных на построении аппроксимирующих функций распределения, учитывающих возможность наличия процессов самоорганизации для событий, описываемых новостными лентами, и памяти (на возможность чего указывают результаты R/S анализа процессов изменения структуры новостных кластеров) о предыдущих состояниях. При этом прогнозируемое событие может быть сконструировано из уже произошедших с использованием полученных теоретических аппроксимирующих функций распределения. Полученные при этом результаты могут быть использованы в аналитических и прогностических целях, например, для определения вероятности повышения в будущем террористической активности.

Третья глава диссертационного исследования посвящена разработке математической модели прогнозирования динамики новостных текстов.

Были выделены основные предположения для создания модели, учитывающие самоорганизацию, «частичную память», нечеткость и неопределенность, самоподобие.

Модель прогнозирования динамики новостных лент разрабатывалась следующим образом:

- Для описания схемы переходов между возможными состояниями, описывающими некоторый процесс в информационном пространстве, был использован аппарат теории вероятностей, что позволило получить алгебраическое уравнение, описывающее вероятности соответствующих переходов.
- Для разложения в ряд Тейлора членов вероятностного алгебраического уравнения, был использован аппарат классического математического анализа, что позволило получить в конечном итоге для описания поведения системы (функции зависимости плотности вероятности от времени) дифференциальное уравнение второго порядка.
- Применение системного анализа позволило учесть при описании процессов переходов между состояниями наличие в описываемых системах памяти и возможности самоорганизации, а также найти граничные и начальные условия для формулировки краевой задачи.
- Для решения краевой задачи, было использовано операционное исчисление и преобразования Лапласа.
- При анализе полученного решения краевой задачи и создании выводов по результатам моделирования были использованы методы системного анализа.

В третьей главе рассматривается модель процессов, происходящих в информационном пространстве. Величина среднего значения проекций векторов x_i в выбранный момент времени может случайным образом увеличиваться на некоторое значение ε, или уменьшаться на некоторое значение ξ (проекции векторов x_i характеризуют положение центров новостных кластеров в информационном пространстве на ось прогнозируемого события новостной ленты). В результате, состояние x_i через некоторое время может оказаться близко к *порогу прогнозируемого события*, по своему значению равному величине вектора X_{bs}.

Вводятся обозначения:

- Множество состояний, описывающих возможность появления на оси прогнозирования некоторого события – X.
- Состояние, наблюдаемое в момент времени $t - x_i$ ($x_i \in X$).
- Временной интервал, за который состояние x_i может измениться – τ.
- Любое значение текущего времени в данном случае: t=h τ, где h – номер шага перехода между состояниями, h=0,1,2,3,…,N.
- Глубина памяти (число шагов изменения состояния системы) – m.

После перехода с шага h на шаг h+1 текущее состояние x_i может уменьшиться на некоторую величину ε или увеличиться на величину ξ, и соответственно, стать равным $x_i - \varepsilon$ или $x_i + \xi$. Величины ξ и ε являются параметрами моделируемых процессов и принадлежат области определения x_i. Дополнительно на $x_i - \varepsilon$ и $x_i + \xi$ накладываются ограничения:

$x_i + \xi \leq L_1$, где L_1 – верхняя граница множества X,

$x_i - \varepsilon \geq L_2$, где L_2 – нижняя граница множества X.

ε и ξ, в самом простом случае, для любого шага h являются постоянными величинами.

Вводится понятие «вероятности нахождения информационного пространства в том или ином состоянии». Допустим, что после известного числа шагов h про описываемую нами систему можно сказать, что:

- P(x, h) – вероятность того, что она находится в состоянии x;
- P(x–ε, h) – вероятность того, что она находится в состоянии (x–ε);
- P(x+ξ, h) – вероятность того, что она находится в состоянии (x+ξ).

После каждого шага состояние x_i может изменяться на величину ε и ξ (далее индекс i для краткости можно опустить).

Вероятность того, что на (h+1) шаге система (или процесс) окажется в состоянии x равна (см. рис. 1):

$$P(x, h+1)= P(x–\varepsilon, h)+ P(x+\xi, h)– P(x, h) \quad (1)$$

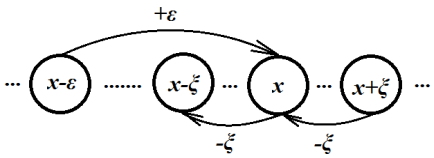

Рисунок 1 – Возможные переходы между состояниями системы (или процесса) на h+1 шаге

Получаем марковский непрерывный процесс. В нём система не обладает памятью состояний. Тем не менее, в действительности, в системе, которой является наше общество, всегда *остается некая память* о предыдущем её состоянии. Вследствие чего, предлагаемая модель должна учитывать данный факт.

С этой целью вероятности P(x–ε, h), P(x+ξ, h) и P(x, h) выражены через состояния на h–1 шаге и изображены схемы соответствующих переходов (см. рис. 2) аналогично схеме, представленной на рис. 1. Учитывая, что ε и ξ – некоторые постоянные величины, для шага h записывается:

$$P(x–\varepsilon, h)=P(x–2\varepsilon, h–1)+P(x–\varepsilon+\xi, h–1)–P(x–\varepsilon, h–1) \quad (2)$$
$$P(x+\xi, h)=P(x+\xi–\varepsilon, h–1)+P(x+2\xi, h–1)–P(x+\xi, h–1) \quad (3)$$
$$P(x, h)=P(x–\varepsilon, h–1)+P(x+\xi, h–1)–P(x, h–1) \quad (4)$$

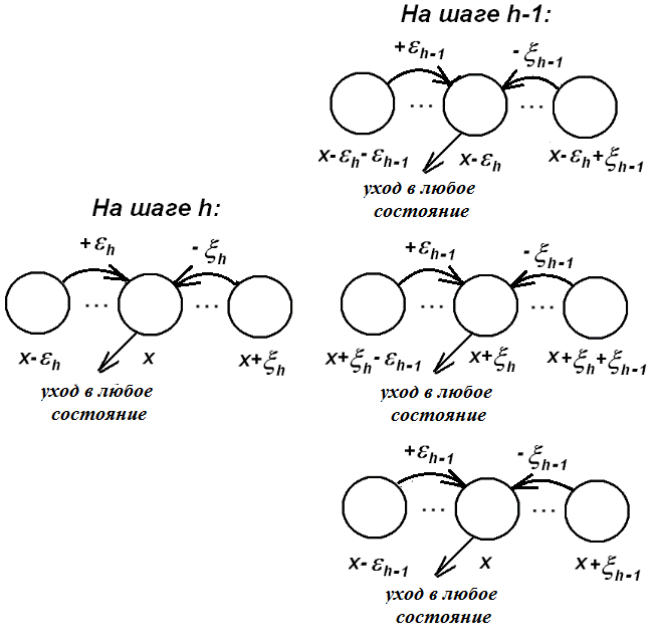

Рисунок 2 – Схема возможных переходов на h–1 шаге
для определения вероятностей P(x–ε, h), P(x+ξ, h) и P(x, h)

Подставив (2), (3) и (4) в уравнение (1), получили:

$$P(x, h+1)=\{P(x–2\varepsilon, h–1)+P(x–\varepsilon+\xi, h–1)–P(x–\varepsilon, h–1)\}+\{P(x+\xi–\varepsilon, h–1)+$$
$$+P(x+2\xi, h–1)–P(x+\xi, h–1)\} –P(x–\varepsilon, h–1) –P(x+\xi, h–1)+P(x, h–1) \quad (5)$$

Далее (5) было преобразовано к виду:

$$\mathrm{P}(x,\mathrm{h}+2)=\{\mathrm{P}(x-2\varepsilon,\mathrm{h})+\mathrm{P}(x-\varepsilon+\xi,\mathrm{h})-\mathrm{P}(x-\varepsilon,\mathrm{h})\}+\{\mathrm{P}(x+\xi-\varepsilon,\mathrm{h})+$$
$$+\mathrm{P}(x+2\xi,\mathrm{h})-\mathrm{P}(x+\xi,\mathrm{h})\}-\mathrm{P}(x-\varepsilon,\mathrm{h})-\mathrm{P}(x+\xi,\mathrm{h})+(x,\mathrm{h}) \quad (6)$$

Выразив члены правой части уравнения (6) через соответствующие вероятности переходов на предыдущем шаге, получили при m=3:

$$\mathrm{P}(x,h+3)=\mathrm{P}(x-3\varepsilon,h)+\mathrm{P}(x+3\xi,h)-\mathrm{P}(x,h)+3\{\mathrm{P}(x-[2\varepsilon-\xi],h)+$$
$$+\mathrm{P}(x-[\varepsilon-2\xi],h)-2\mathrm{P}(x-[\varepsilon-\xi],h)-\mathrm{P}(x-2\varepsilon,h)-\mathrm{P}(x+2\xi,h)+$$
$$+\mathrm{P}(x-\varepsilon,h)+\mathrm{P}(x+\xi,h)\} \quad (7)$$

Были получены и проанализированы формулы для глубины памяти (число шагов изменения состояния системы) m=4,5. Получили для любого значения m рекуррентное выражение вероятности $P(x,h+m)$ того, что для времени $(h+m)$ состояние информационного процесса окажется равным x:

$$\mathrm{P}(x,h+m)=\begin{cases}\sum_{k,q=0}^{m}(-1)^{m-k-q}\dfrac{m!\cdot\mathrm{P}(x-[k\cdot\varepsilon-q\cdot\xi],h)}{k!\cdot q!\,(m-k-q)!}, & \text{при } m-k-q\geq 0\\ 0, & \text{при } m-k-q<0\end{cases}$$

Выполнены соответствующие разложения в ряд Тейлора учитывая, что $t=h\tau$. Здесь t – время процесса, τ – длительность одного шага, h – номер шага. Для этого перешли от h к t:

$$\mathrm{P}(x,t+m\tau)=\begin{cases}\sum_{k,q=0}^{m}(-1)^{m-k-q}\frac{m!\cdot\mathrm{P}(x-[k\cdot\varepsilon-q\cdot\xi],t)}{k!\cdot q!(m-k-q)!}, & \text{при } m-k-q\geq 0\\ 0, & \text{при } m-k-q<0\end{cases} \quad (8)$$

Учитывая производные не выше второго порядка, получили дифференциальное уравнение для изменения вероятности обнаружения информационного процесса в некотором состоянии x, зависящего от времени t и глубины памяти m:

$$\mathrm{P}(x,t)+m\tau\frac{\partial\mathrm{P}(x,t)}{\partial t}+\frac{(m\tau)^2}{2!}\frac{\partial^2\mathrm{P}(x,t)}{\partial t^2}=$$
$$=\sum_{k,q=0}^{m}(-1)^{m-k-q}\frac{m!}{k!\cdot q!(m-k-q)!}\left\{\mathrm{P}(x,t)-[k\cdot\varepsilon-q\cdot\xi]\frac{\partial\mathrm{P}(x,t)}{\partial x}+\frac{[k\cdot\varepsilon-q\cdot\xi]^2}{2!}\frac{\partial^2\mathrm{P}(x,t)}{\partial x^2}\right\} \quad (9)$$

Учитывая, что:

$$\sum_{k,q=0}^{m}(-1)^{m-k-q}\frac{m!}{k!\cdot q!\,(m-k-q)!}=1$$

$$\sum_{k,q=0}^{m}(-1)^{m-k-q}\frac{m!}{k!\cdot q!\,(m-k-q)!}\cdot k=\sum_{k,q=0}^{m}(-1)^{m-k-q}\frac{m!}{k!\cdot q!\,(m-k-q)!}\cdot q=m$$

$$\sum_{k,q=0}^{m}(-1)^{m-k-q}\frac{m!\cdot k^2}{k!\cdot q!\,(m-k-q)!}=\sum_{k,q=0}^{m}(-1)^{m-k-q}\frac{m!\cdot q^2}{k!\cdot q!\,(m-k-q)!}=m^2$$

$$\sum_{k,q=0}^{m}(-1)^{m-k-q}\frac{m!\cdot k\cdot q}{k!\cdot q!\,(m-k-q)!}=m(m-1),$$

получили:

$$m\tau\frac{\partial\mathrm{P}(x,t)}{\partial t}+\frac{(m\tau)^2}{2!}\frac{\partial^2\mathrm{P}(x,t)}{\partial t^2}=$$
$$=\frac{1}{2}\{m^2\varepsilon^2-2m(m-1)\varepsilon\xi+m^2\xi^2\}\frac{\partial^2\mathrm{P}(x,t)}{\partial x^2}-m[\varepsilon-\xi]\frac{\partial\mathrm{P}(x,t)}{\partial x} \quad (10)$$

Продифференцировав уравнение (10) по x, и учитывая, что функция $\mathrm{P}(x,t)$ непрерывна, перешли от вероятности $\mathrm{P}(x,t)$ к плотности вероятности $\rho(x,t)=\frac{\partial\mathrm{P}(x,t)}{\partial x}$ обнаружения информационного процесса в некотором состоянии x в зависимости от величины времени t и глубины памяти m (при m от 1 до ∞):

$$\frac{\partial\rho(x,t)}{\partial t}=\frac{m\varepsilon^2-2(m-1)\varepsilon\xi+m\xi^2}{2\tau}\cdot\frac{\partial^2\rho(x,t)}{\partial x^2}-\frac{\varepsilon-\xi}{\tau}\cdot\frac{\partial\rho(x,t)}{\partial x}-\frac{m\tau}{2}\cdot\frac{\partial^2\rho(x,t)}{\partial t^2} \quad (11)$$

Замечание. При m=1 уравнение (11) описывает марковские процессы. При $m\geq 2$ уравнение (11) позволяет описывать немарковские процессы с различной глубиной памяти m.

Член уравнения вида $\frac{\partial\rho(x,t)}{\partial x}$ описывает упорядоченный переход в одно из двух состояний: либо когда оно увеличивается ($\varepsilon > \xi$), либо когда уменьшается ($\varepsilon < \xi$); $\frac{\partial^2\rho(x,t)}{\partial x^2}$ описывает случайное изменение состояния (*неопределенность изменения*); $\frac{\partial\rho(x,t)}{\partial \mathrm{t}}$ показывает скорость общего изменения состояния системы с течением времени; $\frac{\partial^2\rho(x,t)}{\partial \mathrm{t}^2}$ описывает процесс, в результате которого состояния сами становятся источниками появления других состояний (*самоорганизация* за счет ускорения случайных ($\frac{\partial^2\rho(x,t)}{\partial x^2}$) и упорядоченных ($\frac{\partial\rho(x,t)}{\partial x}$) переходов).

Учитывая область применимости модели, необходимо в уравнении (11) учесть ограничение, накладываемое на коэффициент $(m\varepsilon^2 - 2(\mathrm{m}-1)\varepsilon\xi + \mathrm{m}\xi^2)/2\tau$. Этот коэффициент находится перед второй производной по x, учитывающей возможность случайного изменения состояния системы. Если $(m\varepsilon^2 - 2(\mathrm{m}-1)\varepsilon\xi + \mathrm{m}\xi^2) < (l - x_0)^2$, то система за один шаг перейдет через порог достижения события. В связи с тем, что переход через порог достижения события (l) из начального состояния x_0 не может произойти быстрее, чем за время одного шага τ, должно выполняться условие: $(m\varepsilon^2 - 2(\mathrm{m}-1)\varepsilon\xi + \mathrm{m}\xi^2) \geq (l - x_0)^2$.

Сформулирована граничная задача. Её решение будет описывать процесс перехода между состояниями в информационном пространстве.

Первое граничное условие. При выборе первого граничного условия будем исходить из следующих соображений: состояние x=0 определяет полное отсутствие любых процессов, с соответствующими им измеряемыми параметрами, которые протекают в информационном пространстве. Вероятность обнаружить такое состояние системы может быть не равна 0 (но должна быть близка к 0). А плотность вероятности, которая определяет поток в состоянии $x = 0$, нужно взять равной 0, т.к. состояния системы не могут иметь отрицательные значения (здесь реализуется условие отражения). Таким образом:

$$\rho(x,t)_{x=0} = 0 \qquad \text{(a)}$$

Второе граничное условие. Ограничим область возможных состояний информационной системы некоторой величиной L и выберем второе краевое условие для состояния x=L=1 (косинусная метрика, используемая при расчётах, не может быть больше единицы). Вероятность обнаружить такое состояние с течением времени будет отлична от нуля. Однако плотность вероятности, определяющую поток в состоянии x=L=1, необходимо положить равной нулю (состояния системы не могут превышать максимально возможную величину (реализуется условие отражения от границы)):

$$\rho(x,t)_{x=L} = 0 \qquad \text{(b)}$$

Уравнение (11) содержит вторую производную по времени. Для формулировки краевой задачи необходимо задать два начальных условия. Так как в момент времени $t = 0$ состояние системы может достичь некоторого значения x_0, первое начальное условие можно задать в виде:

$$\rho(x, \mathrm{t} = 0) = \delta(x - x_0) = \begin{cases} \int \delta(x - x_0)dx = 1, x = x_0 \\ 0, \qquad\qquad x \neq x_0 \end{cases}$$

Таким образом, из задания начального условия следует, что решение для $\rho(x,t)$ разбивается на две области при $x > x_0$ и при $x \leq x_0$. Наличие δ – функции приводит к тому, что решение, оставаясь непрерывным в точке $x = x_0$, испытывает в ней разрыв производной.

Поскольку в уравнении (11) имеется вторая производная по t, то для его решения необходимо задать второе начальное условие, которое задает скорость изменения плотности вероятности для любого значения амплитуды. Это условие не является столь очевидным, как первое, но в данном случае можно использовать непрерывность функции для любого момента времени.

Второе начальное условие:

$$\left.\frac{\partial\rho(x,t)}{\partial t}\right|_{t=0} = 0$$

Это условие нулевой скорости изменения плотности вероятности любого значения

амплитуды, для интервала времени $t = 0$ (хотя возможны и иные условия, вид которых определяется логикой рассматриваемого процесса).

Используя методы операционного исчисления для плотности вероятности $\rho_1(x,t)$ и $\rho_2(x,t)$ обнаружения состояния системы в одном из значений на отрезке от 0 до L получили следующую систему уравнений.

При $x \geq x_0$

$$\rho_1(x,t) = -\frac{2}{L} e^{-\frac{t}{m\tau}} e^{\frac{(x-x_0)(\varepsilon-\xi)}{m\varepsilon^2-2(m-1)\varepsilon\xi+m\xi^2}} \sum_{n=1}^{\infty} A(t,n|m) \frac{sin\left(\pi n\frac{x_0}{L}\right) sin\left(\pi n\frac{L-x}{L}\right)}{\cos(\pi n)} \quad (12a)$$

При $x < x_0$

$$\rho_2(x,t) = -\frac{2}{L} e^{-\frac{t}{m\tau}} e^{\frac{(x-x_0)(\varepsilon-\xi)}{m\varepsilon^2-2(m-1)\varepsilon\xi+m\xi^2}} \sum_{n=1}^{\infty} A(t,n|m) \frac{sin\left(\pi n\frac{L-x_0}{L}\right) sin\left(\pi n\frac{x}{L}\right)}{\cos(\pi n)} \quad (12b)$$

$$A(t,n|m) = ch\left(\frac{t}{\tau}\sqrt{\frac{1}{m^2}\left\{\frac{2\varepsilon\xi}{m\varepsilon^2-2(m-1)\varepsilon\xi+m\xi^2}\right\} - \frac{\pi^2 n^2 (m\varepsilon^2-2(m-1)\varepsilon\xi+m\xi^2)}{mL^2}}\right)$$

Если реализация прогнозируемого события зависит от увеличения величины начального состояния системы x_0, интеграл $\mathrm{P}(l,t)$:

$$P(l,t) = \int_0^{x_0} \rho_2(x,t)dx + \int_{x_0}^{l} \rho_1(x,t)dx \quad (13)$$

задает вероятность того, что к моменту времени t состояние системы находится на отрезке от *0* до *l* (*l*=X_{bs}). Это означает, что *порог события l* не будет достигнут.

В качестве порога реализации можно взять среднее значение косинуса угла между центроидами кластеров и вектором текста прогнозируемого события.

Вероятность $Q(l,t)$ того, что *порог события l* к моменту времени t окажется достигнутым или превзойденным, вычисляется по следующей формуле:

$$Q(l,t) = 1 - \mathrm{P}(l,t) \quad (14)$$

При любых t и x значения $\rho_1(x,t)$ и $\rho_2(x,t)$ не являются отрицательными, а для функции $Q(l,t)$ при $t\to\infty$ выполняется условие $Q(l,t)\to 1$ и $\mathrm{P}(l,t)\to 0$.

Согласно созданной модели, стохастическая динамика описывается изменением состояния системы за счет параметров ε и ξ.

Глава 4 посвящена экспериментальной проверке, созданной в третьей главе, модели прогнозирования динамики новостных лент.

Предложена методика применения модели прогнозирования динамики новостных лент, основанная на математической модели, созданной в третьей главе данного диссертационного исследования.

Экспериментальная проверка модели основана на том, что можно взять уже реализовавшийся за какой-то интервал времени временной ряд, описывающий динамику какого-либо типа событий в новостной ленте. Затем на основе анализа первоначальной части этого ряда определить величины параметров ξ и ε. Далее в качестве прогнозируемого события можно взять текстовое описание новостного события, которое входит в последующую часть временного ряда. И рассчитать для него зависимость от времени вероятности реализации, а затем сопоставить полученные данные с наблюдаемым временем реализации.

1) Из новостей, собранных за 2016 год создали вектора и разделили их на W кластеров по различным темам (в нашем случае W=300 т.е. оказалось 299 тематических + один кластер, в который попали все, не вошедшие в 299 тематических кластера, тексты). Далее, каждый из W кластеров разделили на 365 подгрупп векторов текстов по дням появления новостей. Если в данный день тематических новостей не оказалось, то в дневной подгруппе данного кластера будет пустое множество векторов. Таким образом, в каждом из кластеров, события новостной ленты за 2016 год образуют временные ряды, из которых будут определены параметры модели.

2) Для проверки модели и определения её параметров использовали текстовое описание тематического события, произошедшего в 2017 году (опубликованный текст с датой реализации события, описанного в новости). Создали его вектор X_{bs}.

3) Для каждого дня 2016 года, внутри каждой дневной подгруппы векторов каждого кластера,

определили координаты центроидов: $C_j(t) = \{c(t)_{1,j}, c(t)_{2,j}, \cdots c(t)_{k,j}, \cdots c(t)_{M,j}\}$, где $c(t)_{k,j}$ – среднее арифметическое координат входящих в подгруппу векторов (для данного дня без накопления за предыдущие периоды) в момент времени t, где j принимает значения от 1 до W (т.е. для каждого дня получили W центроидов). Если в данный день тематических новостей не оказалось, то в дневной подгруппе данного кластера будет пустое множество векторов и центроид тоже образует пустое множество.

4) Внутри каждого из кластеров W вычислили для каждого момента времени t =1,2,3,…..365 (каждого дня) величины косинусов углов между векторами дневных центроидов $C_j(t)$ и вектором новости X_{bs} текстового описания прогнозируемого события (обозначили эти косинусы, как $S_j(t) = cos\{C_j(t); X_{bs}\}$). Если в данный день новостей нет, то косинусная метрика равна пустому множеству.

5) Выбрали отличные от пустого множества значения косинусной метрики и соответствующие им дни года. Далее провели следующую процедуру по всем не нулевым значениям косинусной метрики до конца года: берем первое ($S_j(t_1)$) и второе ($S_j(t_2)$) значение и находим разность между вторым и первым ($\Delta S_j(t_2 - t_1) = S_j(t_2) - S_j(t_1)$), а затем делим её на интервал времени ($t_2 - t_1$) в днях между вторым и первым отличными от пустого множества значениями косинусной метрики. Таким образом, находим приведенное к одному дню (τ=1) отклонение ($\Delta_j(t_2 - t_1) = \frac{S_j(t_2) - S_j(t_1)}{t_2 - t_1}$ – оно может быть, как положительным, так и отрицательным). Затем берем третье ($S_j(t_3)$) не нулевое значение косинусной метрики, вычитаем из него второе ($S_j(t_2)$) и полученную разность ($\Delta S_j(t_3 - t_2) = S_j(t_3) - S_j(t_2)$), делим на интервал времени ($t_3 - t_2$) в днях между третьим и вторым значением косинусной метрики. Таким образом, опять находим приведенное к одному дню (τ=1) отклонение ($\Delta_j(t_3 - t_2) = \frac{S_j(t_3) - S_j(t_2)}{t_3 - t_2}$ – оно может быть, как положительным, так и отрицательным).

6) Сортируем все отклонения на две группы: $\Delta_j(\Delta t) < 0$ и $\Delta_j(\Delta t) > 0$ и по каждой из них находим средние значения (сумма $\Delta_j(\Delta t)$ деленная на их число). Среднее значение для косинуса отклонения по группе $\overline{\Delta_j(\Delta t)} < 0$ приняли за величину тренда уменьшения ξ, а по группе $\overline{\Delta_j(\Delta t)} > 0$ – за величину тренда увеличения ε.

7) Последнее среднее значение косинусной метрики в конце года (без учета числа пустых множеств) приняли за начальное состояние системы x_0.

В качестве прогнозируемых событий случайным образом были выбраны 10 новостных сообщений (более подробно описано в диссертации), произошедших в 2017 г. Используя модель, описанную в третьей главе данной диссертации, и, созданные во время предобработки данных, текстовые кластеры (300 кластеров), для каждого из 10 прогнозируемых событий новостной ленты были определены значения параметров модели ξ, ε и x_0 (при нахождении ξ и ε было использовано τ=1 день). Далее, меняя глубину памяти (m=2,3,…) и, используя уравнения (12) – (14), была рассчитана зависимость от времени вероятности реализации каждого из прогнозируемых событий, а затем сопоставлены полученные данные с наблюдаемым временем реализации.

Для оценки величины косинусной меры (l) порога реализации события, присутствующей в формуле (13) рассмотрен текстовый пример, в котором два документа S_1 и S_2 имеют очень близкие смысловые значения:

S_1 = «купить книжный шкаф со скидкой»,

S_2 = «недорого купить книжный шкаф с бесплатной доставкой».

Составлена таблица нормализованных (лемматизированных) слов данных предложений:

Таблица 1 – Нормализованные слова предложений тестового примера

	недорог	купить	книжный	шкаф	бесплатно	доставка	скидка
S_1	0	1	1	1	0	0	1
S_2	1	1	1	1	1	1	0

Вычисление косинусной метрики дает значение равное 0,61. В данном случае мы

рассмотрели очень короткие тексты, имеющие большое смысловое сходство. По мере увеличение длины текстов значение косинусной метрики будет существенно уменьшаться, хотя их смысловое значение будет оставаться очень близким, поэтому можно принять, что l=0,5.

Крупные черные точки на рисунках 3 – 5 соответствуют моментам времени фактической реализации событий.

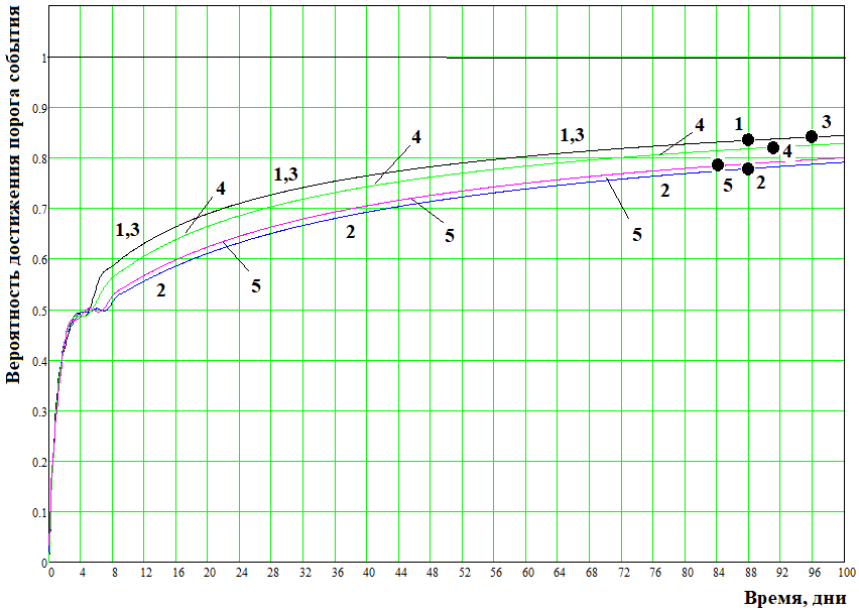

Рисунок 3 – Результаты моделирования преодоления порога событий для 5 новостей (m=2, l=0,5)

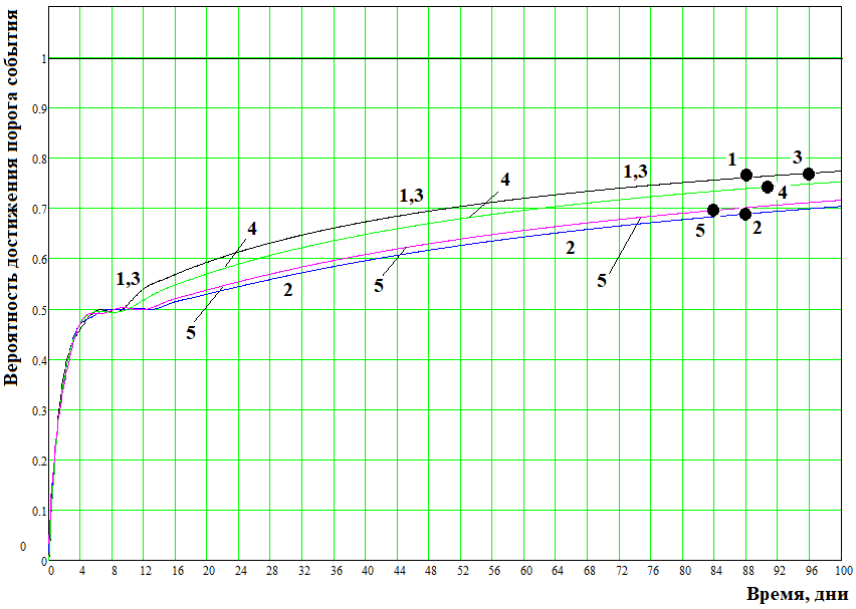

Рисунок 4 – Результаты моделирования преодоления порога событий для 5 новостей (m=3, l=0,5)

Приведенные результаты показывают, что при учете памяти для немарковских процессов с ростом её глубины, величина вероятности реализации событий, которой оно реально соответствовало, уменьшается. При m=2 величина вероятности лежит в диапазоне от 0,78 до 0,85 (см. рис. 3), а при m=3 в диапазоне от 0,70 до 0,77 (см. рис. 4). Ещё больший учет числа шагов будет приводить к ещё более медленному росту вероятности реализации события с течением времени. При m=1 (немарковский процесс, память не учитывается) величина вероятности лежит в диапазоне от 0,90 до 0,92.

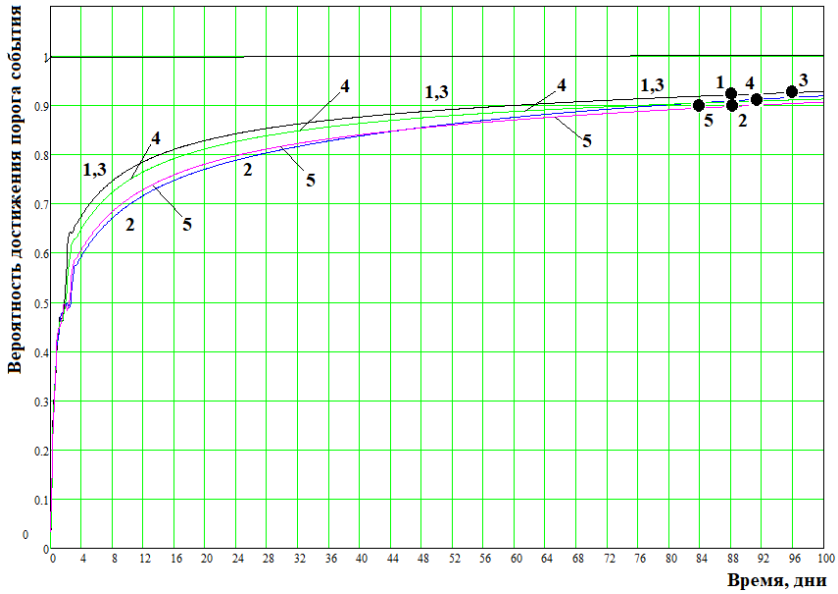

Рисунок 5 – Результаты моделирования преодоления порога событий для 5 новостей (m=1, *l*=0,5)

Так же была проведена проверка разработанной модели на способность прогнозирования фиктивной новости (то, чего не может происходить на самом деле). В качестве примера, взяли небольшой отрывок из русской народной сказки про Колобка. Моделирование динамики вероятности реализации такой новости с течением времени для разработанной модели дает оценку времени его реализации (при величине вероятности 0,8) равную около 50000 дней (примерно 137 лет), что значительно превышает жизнь одного поколения и является маловероятным.

Предложена методика оценки адекватности прогностической модели по экспериментальным данным. Предлагается провести оценочный анализ и сопоставление вероятностей реализации прогнозируемого ($P_{прог}$) и случайного событий ($P_{случ}$). На основе расчетов (см. таб. 2) был сделан вывод о том, что разработанная модель может быть использована для прогнозирования. Относительная точность прогнозирования в зависимости от учета глубины памяти будет выше 60%.

Таблица 2 – Величины точности для пяти новостей

№ новости	Глубина памяти m						Простая диффузионная модель	
	m=1		m=2		m=3			
	Точность Υ, %	Квадрат отклонения σ^2, %	Точность Υ, %	Квадрат отклонения σ^2, %	Точность Υ, %	Квадрат отклонения σ^2, %	Точность Υ, %	Квадрат отклонения σ^2, %
1	87,7	1,00	78,1	9,6	68,5	17,6	79,5	16,0
2	86,3	0,16	69,9	26,0	57,5	46,2	74,0	2,3
3	87,5	0,64	79,2	17,6	69,1	23,0	79,2	13,7
4	86,5	0,04	75,7	0,5	66,2	3,6	73,0	6,3
5	85,3	1,96	72,0	9,0	60,0	18,5	72,0	12,3
Среднее значение, %	$\overline{\Upsilon_1}$=86,7	$\overline{\sigma_1}$=±0,9	$\overline{\Upsilon_2}$=75,0	$\overline{\sigma_2}$=±3,5	$\overline{\Upsilon_3}$=64,3	$\overline{\sigma_3}$=±4,7	$\overline{\Upsilon}$=75,5	$\overline{\sigma}$=±3,2

По полученным в результате экспериментов данным, были сделаны выводы о том, что разработанная модель прогнозирования появления события в сообщении новостной ленты на основе описания процессов с реализацией памяти и самоорганизацией является адекватной и не противоречивой (все новостные события, использованные для проверки модели, в зависимости от того, какая глубина памяти учитывается, могут реализоваться при высоких значениях

вероятности или, если они являются фиктивными, то могут реализоваться только за неприемлемо большое время).

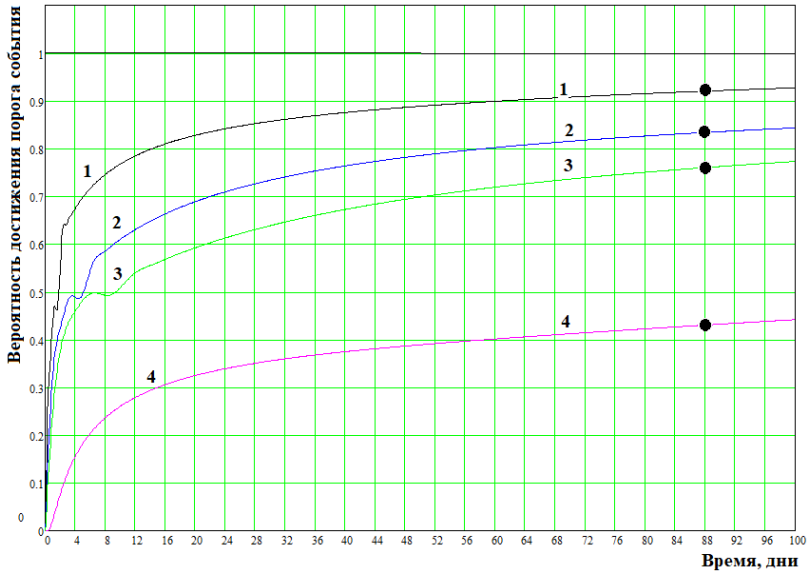

Рисунок 6 – Результаты моделирования преодоления порога событий (при l=0,5) для новости (с использованием разработанной модели, учитывающей самоорганизацию и память: m=1 – кривая 1, m=2 – кривая 2, m=3 – кривая 3, и простой диффузионной модели – кривая 4)

Для проверки возможностей разработанной модели было проведено её сравнение с простой диффузионной моделью, которое показало, что значение вероятностей в моменты времени, когда реально реализовываются, использованные для проверки, события, составляет для простой диффузионной модели менее 50%, а разработанная модель, учитывающая самоорганизацию дает, в зависимости от глубины памяти значения от 76% до 93%. Это является более приемлемым и логичным для прогнозирования динамики события, чем простая диффузионная модель. Моделирование динамики вероятности реализации новости про «Колобка» с течением времени для простой диффузионной модели дает оценку времени его реализации (при величине вероятности 0,8 около 90000 дней (240 лет).

Анализ разработанной модели прогнозирования и её сравнение с простой диффузионной моделью подтверждает возможность прогнозирования появления события в сообщении новостной ленты исходя из его текстового описания, векторизации и нахождения значения косинуса угла между данным вектором и центроидами различных информационных кластеров. Изменение данного косинуса с течением времени можно рассматривать как блуждания точки на отрезке [0,1], который содержит в l ловушку, куда может с течением времени попасть блуждающая точка. Результаты моделирования зависимости от времени вероятности реализации событий с экспериментально определенными наборами значений параметров разработанной модели не являются противоречивыми с точки зрения поведения вероятности (при больших временах вероятности асимптотически стремятся к единице).

Разработанная модель не основывается на статистических характеристиках процессов с заранее предполагаемым законом распределения и учитывает основные свойства реализации событий в информационном пространстве, такие как: неопределенность во времени их проявления, стохастичность, наличие памяти в системе в которой происходит событие, самоорганизация информации. Однако, если в новостном пространстве присутствует память, то глубина этой памяти является не очень большой.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

1. Разработана стохастическая модель прогнозирования динамики новостной ленты:
   – построены разностные схемы для вероятностей переходов между состояниями информационной системы, описывающими эволюцию рассматриваемого процесса с течением времени. При описании процессов перехода между состояниями учтены

возможность самоорганизации и наличие памяти;
- получено алгебраическое уравнение, описывающее условные вероятности соответствующих переходов между возможными состояниями информационной системы;
- получено дифференциальное уравнение второго порядка, описывающее поведение системы (функции зависимости плотности вероятности от времени);
- сформулированы граничные и начальные условия для краевой задачи, решение которой будет описывать процесс перехода между состояниями в информационном пространстве, решена краевая задача.

2. Разработана методика применения созданной модели для прогнозирования появления события в сообщении новостной ленты на основе его формирования из существующих событий новостной ленты.
3. Проведена экспериментальная проверка модели прогнозирования динамики новостных лент. Для этого был решён ряд дополнительных задач:
   - выделены основные характеристики новостных событий, которые нужно учесть при создании математической модели прогнозирования динамики новостных текстов;
   - выполнена лингвистическая обработка текстов новостных сообщений;
   - формализовано математическое представление процессов реального мира, выраженное через их новостное описание, используя тексты на естественных языках для создания их информационных образов;
   - кластеризованны по смысловым группам векторы новостного пространства, выполнено уточнение кластеризации;
   - сформированы временные ряды изменения структуры полученных кластеров с течением времени;
   - разработаны методы анализа временных рядов, описывающих изменение структуры новостных текстовых кластеров заданной тематики от времени с помощью теории самоподобия (R/S анализа).
4. Разработана методика оценки адекватности прогностической модели по экспериментальным результатам.
5. Внедрены научные положения и результаты работы в учебный процесс и в практические разработки коммерческой компании ООО «РУСНЕФТЕТРЕЙД», ООО НАУЧНО-ТЕХНИЧЕСКИЙ ЦЕНТР «ЭССЗ».

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Основные результаты диссертационной работы изложены в **13** публикациях, приведенных в списке литературы, и докладывались на следующих научно-практических конференциях:

Список публикаций в журналах из перечня ВАК:

1. **Новикова О.А.** Моделирование и прогнозирование динамики событий в новостных лентах на основе простой диффузионной модели // Cloud of Science. 2020. – Т. 7. – № 3. – С. 619–643.
2. **Новикова О.А.**, Андрианова Е.Г. Роль методов интеллектуального анализа текста в автоматизации прогнозирования рынка ценных бумаг // Cloud of Science. 2018. – Т. 5. – № 1. – С. 196–211.
3. Отраднов К.К., Жуков Д.О., **Новикова О.А.** Модель кластеризации слабоструктурированных текстовых данных // Современные информационные технологии и ИТ-образование. 2017. – Т. 13. – № 3. – С. 100–115.
4. Жуков Д.О., **Новикова О.А.**, Алёшкин А.С. Возможность использования почти-периодических функций, вейвлет анализа и теории самоподобия Хёрста для прогнозирования новостных событий в информационном пространстве // Современные информационные технологии и ИТ-образование. 2017. – Т. 13. – № 1. – С. 9–18.
5. **Новикова О.А.**, Кошкин Д.Е. Энтропийная оценка качества автоматического разбиения

категорийных данных алгоритмом Fuzzy C–means // Информатизация образования и науки. – 2013. – №1(17). – С. 113–121.

Список публикаций в трудах международных конференций, входящих в базы SCOPUS и Web of Science:

1. Zhukov D.O., Zamyshlyaev A.M., **Novikova O.A.** Model of forecasting the social news events on the basis of stochastic dynamics methods. ITM Web of Conferences. – 2017. – Т. 10. – С. 2009. – ISSN: 2271-2097, WOS: 000406704600017.
2. Zhukov D.O., **Novikova O.A.**, Otradnov K.K. Methods of analysis of news events in the information space based on the use of almost – periodic functions, wavelet transforms and Hurst's self-similarity / Proceeding The 7th International Conference on Information Communication and Management ICICM'17, August 28-30. Moscow, Russian Federation, ACM. – 2017. –P. 95-103. – ISBN: 978-1-4503-5279-6.
3. Sigov, A., Zhukov, D., **Novikova, O.** Modelling of memory realization processes and the implementation of information self-organization in forecasting the new's events using arrays of natural language texts / Proceeding the 1st International Scientific Conference Convergent Cognitive Information Technologies, Convergent 2016; Moscow; Russian Federation; 25 November 2016 through 26 November 2016; Code 125487, CEUR Workshop Proceedings Volume 1763. – 2016. – P. 42–55. – ISSN: 16130073.

Список иных публикаций:

1. **Новикова О.А.** Алгоритм анализа модели стохастической динамики формирования новостных событий // Компьютерные науки и информационные технологии: Материалы Междунар. науч. конф. – Саратов: Издат. центр «Наука». – 2018. –С. 290–296.
2. Сигов А.С., Жуков Д.О., **Новикова О.А.** Моделирование процессов реализации памяти и самоорганизации информации при прогнозировании новостных событий с использованием массивов естественно-языковых текстов // Современные информационные технологии и ИТ – образование. 2016. –Т. 12. – № 1. – С. 42 – 55.
3. **Новикова О.А.** Обзор методов интеллектуального анализа данных применительно к задаче автоматизированного построения онтологий//Современные вопросы науки и образования-XXI век: сб. науч. Трудов по материалам Международной заочной научно-практической конференции 29 февраля 2012 г.: в 7 частях. Часть 3; Мин. Образования и науки Рос. Федерации. Тамбов: Изд-во ТРОО «Бизнес-Наука-Общество». –2012. –164 с.
4. Кошкин Д.Е., **Новикова О.А.** Уточнение кластеризации категорийных данных через оценку энтропии результирующих кластеров // Всероссийский конкурс научно-исследовательских работ студентов и аспирантов в области информатики и информационных технологий: сборник научных работ: в 3 т. –Белгород: ИД «Белгород». –2012. –Т. 3. –С. 167–173.
5. **Новикова О.А.**, Кошкин Д.Е. Уточнение кластеризации категорийных данных через оценку энтропии результирующих кластеров // Итоги диссертационных исследований. Том 2. – Материалы IV Всероссийского конкурса молодых ученых. –М.: РАН. – 2012. –138 с.

Свидетельства

Е.Г. Андрианова, Д.И. Братухин, Л.А. Истратов, **О.А. Новикова**, В.Н. Шумилов Свидетельство о государственной регистрации программы для ЭВМ №2018615544 от 10.05.2018 «Модуль прогнозирования новостных событий на основе анализа спектров информационных процессов». – М.: Роспатент.

Новикова О.А.

РАЗРАБОТКА И ИССЛЕДОВАНИЕ МОДЕЛЕЙ
ПРОГНОЗИРОВАНИЯ ДИНАМИКИ НОВОСТНЫХ ЛЕНТ

Автореферат

Подписано в печать
Формат бумаги 60 × 90 1/16
Отпечатано в Типографии ФАСТКОПИ ИП Сусляков Д.А.,
Свидетельство 77 007140227
129345, г. Москва, ул. Тайнинская, д. 12, к.2, тел. +7(495)643-48-58
Заказ № 114

Объем 1,25 уч.-изд. л.
Тираж 110 экз.