Моделирование и прогнозирование динамики событий в новостных лентах на основе простой диффузионной модели

О. А. Новикова

МИРЭА – Российский технологический университет
119454, Москва, пр-т Вернадского, 78

e-mail: novikova@mirea.ru

Аннотация. Одной из проблем прогнозирования событий новостных лент является разработка моделей, позволяющих работать со слабоструктурированным информационным пространством текстовых документов. В статье описана модель прогнозирования событий в новостных лентах, основанная на стохастической динамике изменений структуры нестационарных временных рядов новостных кластеров (состояний информационного пространства) на основе применения диффузионного приближения. Прогнозирование событий новостной ленты осуществляется исходя из их текстового описания, векторизации и нахождения значения косинуса угла между данным вектором и центроидами различных смысловых кластеров информационного пространства. Изменение значения данного косинуса с течением времени можно рассматривать как блуждания точки на отрезке [0, 1], который содержит ловушку в точке порога реализации события, куда может с течением времени попасть блуждающая точка. При создании модели были рассмотрены схемы вероятностей переходов между состояниями в информационном пространстве. На основании данного подхода было выведено нелинейное дифференциальное уравнение второго порядка, а также сформулирована и решена краевая задача для прогнозирования новостных событий, что позволило получить теоретическую зависимость от времени функции плотности вероятности распределения параметров нестационарных временных рядов, описывающих эволюцию информационного пространства. Результаты моделирования зависимости от времени вероятности реализации событий (с экспериментально определенными, для уже реализовавшихся событий, наборами значений параметров разработанной модели) показывают, что модель является непротиворечивой и адекватной (все новостные события, использованные для проверки модели, в зависимости от того, какая глубина памяти учитывается, реализовываются при высоких значениях вероятности (около 80%), или, если они являются фиктивными, то могут реализоваться только за неприемлемо большое время).

Ключевые слова: прогнозирование событий в новостных лентах, кластеризация новостей, стохастическая динамика, порог новостного события.

1. Введение

За прошедшие десятилетия произошел стремительный рост числа электронных новостных лент, размещенных в сети Интернет и ставших доступными широкому потребителю. Более того, каждый, кто владеет возможностью выхода в мировые информационные сети, сам имеет возможность свободно пополнять мировую копилку знаний. Быстрый рост новостных баз данных привел к востребованности создания инструментов анализа и прогнозирования новостных сообщений, поиску алгоритмов и методов извлечения закономерностей в корпусе новостных лент и прогнозирования вероятности появления в новостной ленте сообщений о значимых событиях в целях упреждающего воздействия на различные виды деятельности и управления их возможными состояниями. К примеру, зная о том, что через определенный промежуток времени могут с большой вероятностью произойти гражданские беспорядки, можно предпринять меры, чтобы их предотвратить или избежать тяжелых последствий.

2. Обзор исследований прогнозирования событий на основе анализа текстов

Разработкой и совершенствованием средств анализа, выявления закономерностей в новостных сообщениях, исследованием особенностей динамики новостных лент занимаются многие российские и зарубежные ученые. Подробно об исследователях и их работах в области интеллектуального анализа текста для прогнозирования рынка ценных бумаг, рассмотрено в [1].

На сегодняшний день большинство работ по прогнозированию на основе анализа текстовых данных сосредоточены вокруг поведения пользователей в соцсетях, на форумах, в чатах. Например, Грал и др. (Gruhl et al.) в своей работе [2] показали, как онлайн общение способно предсказывать продажи книг. А Мишне и Райке (Mishne and Rijke) [3] анализ настроений в блог-постах использовали для предсказания продаж фильмов. В статье Лиу и др. (Liu et al.) [4] описано применение модели вероятностного скрытого семантического анализа PLSA к оценке настроений в блог-постах для предсказаний будущих продаж. А в [5] авторы показали, что поисковые запросы Google способны прогнозировать эпидемиологическое развитие инфекционных заболеваний, а также потребительские предпочтения, расходы. Л. Жао и др. (L. Zhao et al.) [6] показали, как для прогнозирования преступности можно использовать твиты с пространственно-временными метками. В работе к твитам применяется лингвистический анализ, тематическое моделирование (для выделения тематик твитов) для автоматического определения тем, которые затем используются в модели прогнозирования преступности.

Значительно меньше исследований, которые изучают возможность прогнозирования появления новостных событий в информационном пространстве, несмотря на то, что доказано: открытые исходные данные, в том числе новостные, являются суррогатами при прогнозировании различных событий, например, вспышек заболеваний [7], результатов выборов [8, 9] и протестов [10].

Авторы [11] разработали метод, который решает задачу выявления предвестников и прогнозирует события в будущем. По данным из коллекции потоковых новостей (новости взяты из нескольких открытых источников трех стран Латинской Америки) был разработан вложенный подход для прогнозирования значительных общественных событий, протестов. В работе продемонстрированно, как этот подход способен последовательно идентифицировать новостные статьи, считающиеся предвестниками протестов. Сильные стороны предложенного в [11] подхода демонстрирует эмпирическая оценка, которая состоит в фильтрации потенциальных предвестников, в точном прогнозировании характеристик событий гражданских беспорядков.

В работе [12] показана модель для прогнозирования катастроф, стихийных бедствий. Авторы предлагают анализировать исторические данные и, извлекая из них шаблоны событий, связанных с катастрофами, использовать полученные шаблоны в машинном обучении как обучающие выборки, затем прогнозировать предстоящие катастрофы, используя текущие события.

3. Постановка задачи моделирования динамики изменения контента в новостных лентах и методика ее решения

Для прогнозирования появления новых событий с заданной тематикой необходимо создать модель его формирования, например, на основе описания его временного ряда, а затем найти функцию плотности вероятности распределения его параметров. Основной проблемой анализа и моделирования поведения временного ряда событий новостной ленты для получения прогнозов является то, что в любой момент времени имеется только одна реализация процесса (одна статистическая выборка, один образец уже реализовавшегося временного ряда), используя которую нужно создать прогноз для следующих моментов времени.

В существующих методиках анализа, вне зависимости от применяемых инструментов (статистические модели, нейросетевые, модели нечеткой логики и т. д.) нестационарный временной ряд разделяется на отдельные области, в каждой из которых он является квазистационарным, со своей выборочной (для данного участка временного ряда) функцией распределения, а между каждой из областей имеется часть ряда, в которой происходит переходной процесс (разладка). Продолжительность переходного процесса определяется как факторами, характеризующими сме-

ну режима (собственно разладка), так и объемом выборки, который используется для проведения статистического анализа [13]. Параметры выборочной функции распределения могут быть установлены на основе анализа данных, наблюдаемых на интервале времени квазистационарности. На практике необходимо решить две задачи. Первая заключается в определении интервала времени квазистационарности. Вторая — в определении наступления разладки в течение переходного периода, причем с минимальным запаздыванием.

Стационарный временной ряд представляют в виде суммы детерминированной составляющей (трендовой или периодической) и остатка, автокорреляционная функция которого с достаточной точностью близка к нулю, что свидетельствует о близости остатка к «белому шуму». После этого ставится задача о нахождении наиболее близкой статистики (функции распределения), моделирующей поведение остатка.

При исследовании стационарных случайных процессов, согласно теореме Гливенко (о сходимости эмпирической вероятности к теоретическому распределению) [14], чем больше будет учтено наблюдаемых значений, тем точнее могут быть получены теоретические характеристики распределения случайной величины из определенного промежутка. Для нестационарных случайных процессов данное условие, в силу их специфики, не может быть выполнено, что затрудняет возможность использования результатов их анализа для дальнейшего прогнозирования.

Для нестационарных временных рядов индикаторы тех или иных его свойств имеют свой специфический вид, не обобщаемый на ряды другого типа. Например, индикатор линейного тренда не особенно эффективен для рядов с квазипериодическим изменением, как и индикатор нестационарности дисперсии для рядов с квазилинейным трендом. Более того, индикаторы, основанные на некоторых средних характеристиках ряда (например, несколько первых моментов), не образуют базисной системы, по которой можно определить тенденцию локального по времени изменения случайного процесса.

Идентификация состояния нестационарного случайного процесса может быть сформулирована как задача распознавания выборочной функции распределения (ВФР), как принадлежащей определенной генеральной совокупности. Однако, если функция распределения нестационарна, то обучение алгоритма распознавания на прошлых данных часто оказывается несостоятельным. Имеется только одна траектория, которая в силу нестационарности не позволяет использовать большой объем выборки для тестирования тех или иных индикаторов локального поведения временного ряда.

Таким образом, можно сказать, что анализ квазистационарных участков наблюдаемых временных рядов и построение выборочных функций распределения могут оказаться малоэффективными для прогнозирования последующей эволюции.

В настоящее время в применяемых на практике моделях анализа и прогнозирования динамики нестационарных временных рядов (их эволюции), в качестве аппроксимаций распределений чаще всего используются диффузионные уравнения, включая нелинейную диффузию [15], уравнение Лиувилля [16], уравнение Фоккера — Планка [16] и ряд других. Недостатком всех моделей является то, что они не учитывают возможность самоорганизации протекающих процессов и наличие памяти о предыдущих действиях и состояниях.

Использование для моделирования динамики событий новостной ленты существующих методов анализа временных рядов может приводить к существенным ошибкам, это связано с большой изменчивостью их характеристик, а также нелинейностью, нестационарностью, самоорганизацией происходящих процессов и наличием памяти. Поэтому 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. 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.

В некоторых работах описан ряд теоретико-вероятностных подходов к прогнозированию новостных событий. Так, например, в работе [17] изложена модель прогнозирования будущих событий, путем обобщения конкретных наборов последовательностей событий, извлеченных из новостей за 22 года: с 1986 по 2008 годы. Авторы пытаются построить модель, которая учитывает связь между произошедшими историческими событиями и предсказывает будущие события. Авторы предполагают, что события в реальном мире генерируются вероятностной моделью, которая также генерирует новостные сообщения об этих событиях. Сообщения из новостных событий используются для построения модели в форме определения вероятности $P(\mathrm{ev}_j(t+\Delta) \mid \mathrm{ev}_j(t))$ реализации некоторого будущего события ev_j в момент времени $t + \Delta$ и прошедшего в момент времени t события ev_j. Эта вероятность аппроксимирует связь между двумя произошедшими событиями реального мира. Модель показывает, что с вероятностью 18% событие о засухе ev_j происходит после события о потопе ev_j.

Использование текстовых данных и методов машинного обучения для прогнозирования катастроф и стихийных бедствий описано в работе [12]. Авторы собрали по ключевым словам текстовые сообщения о катастрофах из поисковой системы Google. Затем, полученные в результате запросов текстовые документы, обрабатывались методами математической лингвистики, и с использованием обученного байесовского классификатора отсеивались ложные результаты. После сбора данных, проводилась смысловая кластеризация собранных данных. Из ключевых слов, по которым генерировались поисковые запросы, была построена матрица переходов, а из сгруппированных событий построена матрица наблюдений. Затем обе матрицы подавались на вход скрытой марковской модели для составления прогноза. По словам авторов, данный подход позволяет предсказывать будущие события и места, в которых эти события произойдут.

В работе [18] для решения задачи прогнозирования новостных событий авторы изучают временные зависимости в потоках событий и вводят кусочно-постоянную аппроксимацию их интенсивности, применяя Байесовский подход и распределение Пуассона к описанию выборки важности будущих событий. Это позволяет построить нелинейные временные зависимости для предсказания будущих событий с использованием деревьев решений.

Появление с течением времени в новостных лентах описания событий определенного типа, относящихся к заданной тематике (например, упоминания о террористических актах, деятельности тех или иных политических лидеров и т. д.), можно рассматривать как формирование дискретного временного рядя (параметром которого является частота упоминаний данного события в течение суток). Анализ характеристик динамики данного ряда можно использовать для прогнозирования его эволюции и определения отсутствия или наличия в его поведении долговременных зависимостей, а также расчета вероятности реализации событий в течении заданного интервала времени.

Следует отметить, что для формирования временного ряда появления событий в новостной ленте нужно решить важную вспомогательную задачу: необходимо с высокой точностью выделить из всего множества текстовых сообщений новостных лент (сотни тысяч и миллионы), именно те, которые относятся к данной тематике (кластеризация событий по смысловым группам). Обеспечение высокой точности кластеризации гарантирует, что при формировании временного ряда не будет потеряна существенная часть информации, например, по частотам появления данного события, что позволит добиться более точного определения параметров рассматриваемого временного ряда и не окажет влияния на прогноз его эволюции.

Суть предлагаемого подхода, который может быть использован при создании модели конструирования прогноза будущего события из уже произошедших с ис-

пользованием теоретических аппроксимирующих функций распределения состоит в следующем.

1) Возьмем коллекцию (корпус) из N текстовых документов, описывающих события новостной ленты за некоторый период времени с привязкой к датам их появления. Далее, используя лексические и семантические методы математической лингвистики (удаление знаков препинания, стоп-слов, приведение слов к нормальной форме, лемматизации, создание словаря терминов и т. д.) [19–22], создадим с помощью словаря терминов (слова, n-граммы или объекты ассоциативно-семантических классов) размера M, векторное представление множества текстов в информационном пространстве (размерность которого будет R^M). Для повышения точности анализа текстов и дальнейшей кластеризации по смысловым группам можно использовать подходы, основанные на объединении слов, имеющих в текстах схожее значение в ассоциативно-семантические классы, например, с использованием алгоритма **word2vec**.

Каждому документу из множества можно поставить в соответствие вектор $X_i = \{x_{1,i}, x_{2,i}, .., x_{k,i}, \cdot ..., x_{M,i}\}$, где i — принимает значения от 1 до N, а каждый элемент вектора $x_{k,i}$ характеризует нормированную *TFIDF* частоту вхождения k-термина (слова, n-граммы или объекты ассоциативно-семантических классов) из словаря в i-документ коллекции: $TFIDF = TF \times IDF = (n_k / \Sigma_k n_k) \times \log(D/d)$, где n_k — число вхождений k-термина в документ; $\Sigma_k n_k$ — общее количество терминов в документе; D — общее количество документов в коллекции; d — количество документов, в которых встречается данный термин. Использование *TF-IDF* уменьшает вес широкоупотребительных терминов, что является логически обоснованным и в конечном итоге повышает точность кластеризации текстов. Вектора X_i образуют матрицу $N \times N$ (термин — документ): $\| x_{k,j} \|_{N \times M}$.

2) Далее с применением стандартных методов [19–22] и алгоритмов проведем кластеризацию текстовых документов (разделение по смысловым группам), используя их векторное представление.

Для выполнения кластеризации может быть, например, использован обладающий большим числом преимуществ (простота реализации, качество кластеризации и скорость выполнения) и наиболее широко применяемый для этих целей алгоритм k-means, который относится к классу неиерархических, четких, интеграционных алгоритмов. Некоторым недостатком данного алгоритма является необходимость заранее указывать количество кластеров. Однако он имеет высокую скорость работы и сложность $O(j \cdot C \cdot D \cdot \Sigma_k n_k)$, где C — количество кластеров; $\Sigma_k n_k$ — количество значений элементов в векторе; D — количество документов; j — количество

итераций. Цель алгоритма — найти такие центры кластеров, чтобы расстояние между вектором документа кластера и вектором центра кластера (центроидом) было минимально

$$\arg\min_{N_p} \sum_{p=1}^{k} \sum_{y_i \in N_p} d(y_i, \mu_p) = \arg\min_{N_p} \sum_{p=1}^{k} \sum_{y_i \in N_p} \|y_i - \mu_p\|^2,$$

где y_i — документ из кластера; μ_p — центроид кластера p; N_p — набор документов кластера p. Центроид, т. е. арифметический средний вектор всех векторов кластера (или его подгруппы) может быть вычислен следующим образом: $\mu_p = \sum N_p / D_p$, N_p — вектор новости из кластера p; D_p — количество текстов новостей в кластере. В качестве расстояния между векторами может быть использована косинусная метрика (косинус угла между векторами):

$$d(y_i, z_i) = \frac{\sum_{i=1}^{n} (y_i \cdot z_i)}{\sqrt{\sum_{i=1}^{n} y_i^2} \cdot \sqrt{\sum_{i=1}^{n} z_i^1}},$$

y_i — значение координат первого вектора; z_i — значение координат второго вектора (чем больше косинус угла между векторами, тем больше похожи документы). Учитывая, что все элементы векторов являются положительными числами, то $0 \le d(y_i, z_i) \le 1$.

Помимо алгоритма k-means могут быть использованы и другие хорошо зарекомендовавшие себя алгоритмы кластеризации. Например, DbScan, Affinity Propagation», Agglomerative Clustering, BIRCH.

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. Таким образом может быть сформирован временной ряд, описывающий события определенного типа в новостных лентах. Параметрами таких рядов могут, например, являться либо частоты появления сообщений в новостной ленте о данном типе событий, либо положение центроидов кластеров, включающих тексты, описывающие данные события.

Для формирования временных рядов, описывающий события определенного типа в новостных лентах и проведения исследований была собрана с 4 российских новостных сайтов коллекция из 100 тысяч текстовых документов за 2016 г. («Ведомости», «Коммерсантъ», «РосБизнесКонсалтинг», «Новости. Первый канал»). Максимальное количество слов в одном документе составляло: 10404, минимальное — 101. Словарь используемых терминов для создания матрицы «термин – документ» включал 2 570 724 слов и 1 451 828 терминов.

Согласно результатам проведенного исследования и сравнения алгоритмов кластеризации, наилучшее качество и время работы показали неиерархические алгоритмы кластеризации k-means и Affinity Propagation. Алгоритм k-means выделяет кластеры с более общими тематиками, а алгоритм Affinity Propagation выделяет кластеры с подтемами (одна общая тема разделяется на несколько кластеров с более узкими темами). В качестве моделей представления текста наилучший результат показали модели «документ — термин» с *TFIDF* без использования n-грамм и «документ — ассоциативно-семантическая группа».

Из имеющегося новостного корпуса в результате кластеризации было получено 300 кластеров по различным тематикам. Далее каждый из них был сегментирован на 365 подгрупп новостей за сутки (24 часа) без суммирования за предыдущие периоды.

3) 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 Xbs). 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 числовой 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 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* (т. е., по сути, рассматриваются почти случайные блуждания точки на отрезке [0, 1], который содержит в l ловушку, куда может с течением времени попасть блуждающая точка).

Изложенный подход позволяет на основе рассмотрения схем вероятностных переходов между состояниями сформулировать краевую задачу о зависимости вероятности достижения прогнозируемого события от времени и рассмотреть ее решение (получить теоретическую аппроксимирующую функцию распределения).

4. Вывод функции распределения характеристик временного ряда, описывающего динамику контента новостных лент

4.1. Построение разностных схем вероятностей переходов между состояниями в информационном пространстве. Вывод основного вида модели

В качестве меры сходства смысла между двумя текстовыми документами в компьютерной лингвистике часто используют косинусную метрику. Близость величины косинуса к единице говорит о степени совпадения смыслов текстов, а к нулю о различии. Причем само значение косинуса в данном случае всегда будет находиться на отрезке от 0 до 1 ([0, 1]). Обозначим величину текущего среднего значения косинуса угла между вектором текстового описания прогноза и центроидами текстовых кластеров из которых, как мы предполагаем, данное событие может сформироваться, как x_i (состояние информационной системы).

Пусть интервал времени процесса изменения состояний имеет величину τ (бесконечно малая). Предположим, что за интервал времени τ состояние системы может увеличиться на некоторую величину ε (тренд увеличения) или уменьшиться на величину ξ (тренд уменьшения). Обозначим все множество состояний на оси прогнозирования, как X. Состояние, наблюдаемое в момент времени t можно обозначить, как x_i ($x_i \in X$). В конечном счете, состояние системы x_i может окажется вблизи порога прогнозируемого события равного 1 (или заданной в качестве реализации другого значения косинуса угла между центроидом кластера и вектором текста прогнозируемого события).

Запишем значение текущего времени, как $t = h\tau$, где h — номер шага перехода между состояниями (процесс перехода между состояниями становится квазинепрерывным с бесконечно малым временным интервалом τ), $h = 0, 1, \ldots, N$. Текущее состояние x_i на шаге h после перехода на шаге $(h + 1)$ может увеличиваться на некоторую величину ε, или уменьшаться на величину ξ, и, соответственно, оказаться равным $x_i - \varepsilon$ или $x_i + \xi$.

Введем понятие вероятности нахождения информационного пространства в том или ином состоянии. Пусть, после некоторого числа шагов h про описываемую систему можно сказать, что:

- $\mathbf{P}(x - \varepsilon)$ — вероятность того, что она находится в состоянии $(x - \varepsilon)$;
- $\mathbf{P}(x, h)$– вероятность того, что она находится в состоянии x;
- $\mathbf{P}(x + \xi, h)$ — вероятность того, что она находится в состоянии $(x + \xi)$.

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

Вероятность $\mathbf{P}(x, h + 1)$ того, что на следующем $(h + 1)$ шаге система окажется в состоянии x, будет определяться несколькими переходами (см. рис. 1):

$$\mathbf{P}(x,h+1) = \mathbf{P}(x-\varepsilon,h) + \mathbf{P}(x+\xi,h) - \mathbf{P}(x,h). \quad (1)$$

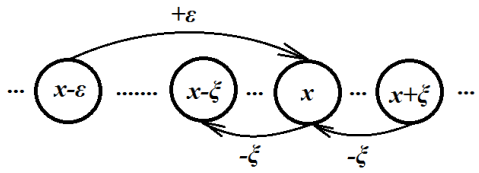

Рисунок 1. Схема возможных переходов между состояниями системы на h+1 шаге

Поясним выражение (1) и представленную на рис. 1 схему. Вероятность перехода в состояние x на шаге h — $\mathbf{P}(x, h + 1)$ определяется суммой вероятностей переходов в это состояние из состояний $(x – \varepsilon)$: $\mathbf{P}(x – \varepsilon, h)$ и $(x + \varepsilon)$: $\mathbf{P}(x + \varepsilon, h)$ в которых находилась система на шаге h за вычетом вероятности перехода $\mathbf{P}(x, h)$ системы из состояния x (в котором она находилась на шаге h) в любое другое состояние на $(h + 1)$ шаге. В данном случае будем считать, что сами переходы осуществляются с вероятностью равной 1.

Проведем соответствующие разложения членов уравнения (1) в ряд Тейлора, и, с учетом производных не выше второго порядка по x и первой производной по времени t, получим следующее дифференциальное уравнение для изменения вероятности обнаружения информационного процесса в некотором состоянии x в зависимости от величины времени:

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

Продифференцировав уравнение (2) по x, перейдем к зависимости плотности вероятности обнаружения информационного процесса в некотором состоянии x в зависимости от величины времени t:

$$\frac{\partial \rho(x,t)}{\partial t} = \frac{\varepsilon^2 + \xi^2}{2\tau}\cdot\frac{\partial^2 \rho(x,t)}{\partial x^2} - \frac{\varepsilon - \xi}{\tau}\cdot\frac{\partial \rho(x,t)}{\partial x}. \quad (3)$$

Данное уравнение можно рассматривать, как уравнение простой диффузионной модели, но еще и учитывающее упорядоченное движение (снос) между состояниями информационного процесса. Член уравнения $\partial\rho(x,t)/\partial t$ показывает скорость изменения состояний системы с течением времени; член уравнения $\partial^2\rho(x,t)/\partial x^2$ учитывает случайные переходы (диффузионные блуждания состояния системы); член уравнения $\partial\rho(x,t)/\partial x$ описывает упорядоченные переходы (тренд или снос), например, либо когда величина состояния увеличивается ($\varepsilon > \xi$), либо когда уменьшается ($\varepsilon < \xi$).

С точки зрения области применимости модели, в уравнении (3) необходимо учесть ограничение, накладываемое на коэффициент $(\varepsilon^2+\xi^2)/(2\tau)$ перед второй производной по x, которая учитывает возможность случайного изменения состояния. Должно выполняться условие $(\varepsilon^2+\xi^2)<(l-x_0)^2$, смысл которого заключается в том, что переход из начального состояния x_0 через порог достижения события (l) не может произойти быстрее, чем за время одного шага τ. Если $(\varepsilon^2+\xi^2)\geq(l-x_0)^2$, то система переходит через порог достижения события за один шаг.

4.2. Формулировка и решение краевой задачи при прогнозировании новостных событий в информационном пространстве для систем с реализацией памяти и самоорганизацией

Считая функцию **P**(x, t) непрерывной, можно перейти от вероятности **P**(x, t) (уравнение (3)) к плотности вероятности $\rho(x,t)=\partial P(x,t)/\partial x$ и сформулировать краевую задачу, решение которой и будет описывать процесс перехода между состояниями в информационном пространстве.

Первое краевое условие. Первое краевое условие выберем для состояния $x = 0$. Вероятность обнаружить такое состояние с течением времени может быть отлична от 0, однако плотность вероятности, определяющая поток в состоянии $x = 0$, необходимо положить равной 0 (состояния системы не могут выходить в область отрицательных значений (реализуется условие отражения, значение косинуса угла в данном случае не может быть отрицательным по определению косинусной метрики для векторов текстов)), т. е.:

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

Второе краевое условие. Ограничим область возможных состояний информационной системы величиной L (косинусная метрика не может быть больше 1) и выберем второе краевое условие для состояния $x = L = 1$. Вероятность обнаружить такое состояние с течением времени будет отлична от 0. Однако плотность вероятности, определяющую поток в состоянии $x = L = 1$ необходимо положить равной 0 (состояния системы не могут выходить в область значений больше, чем максимально возможная величина (реализуется условие отражения от границы)), т. е.:

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

Поскольку в момент времени $t = 0$ состояние системы уже может быть равно некоторому значению x_0, то начальное условие зададим в виде:

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

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, можно получить следующую систему уравнений:

При $x \geq x_0$

$$\rho_1(x,t) = -\frac{2}{L} e^{a_1 \frac{(x-x_0)}{2a}} e^{-\frac{a_1^2 t}{4a}} \sum_{n=1}^{M} \frac{1}{\cos(\pi n)} \sin\left(\pi n \frac{x_0}{L}\right) \sin\left(\pi n \frac{L-x}{L}\right) e^{-\frac{a\pi^2 n^2}{L^2} t}. \quad (4a)$$

При $x < x_0$

$$\rho_2(x,t) = -\frac{2}{L} e^{a_1 \frac{(x-x_0)}{2a}} e^{-\frac{a_1^2 t}{4a}} \sum_{n=1}^{M} \frac{1}{\cos(\pi n)} \sin\left(\pi n \frac{L-x_0}{L}\right) \sin\left(\pi n \frac{x}{L}\right) e^{-\frac{a\pi^2 n^2}{L^2} t}, \quad (4b)$$

где $a = (\varepsilon^2 + \xi^2)/(2\tau)$ и $a_1 = (\varepsilon - \xi)/\tau$.

Если реализация прогнозируемого события связана с увеличением величины исходного состояния системы x_0, то интеграл $\mathbf{P}(l, t)$:

$$\mathbf{P}(l,t) = \int_0^{x_0} \rho_2(x,t)\,dx + \int_{x_0}^{l} \rho_1(x,t)\,dx \quad (5)$$

будет задавать вероятность того, что состояние системы к моменту времени t находится на отрезке от 0 до l, т. е. порог события l не будет достигнут. В качестве порога реализации можно использовать заданное среднее значение косинуса угла между центроидами кластеров и вектором текста прогнозируемого события).

Соответственно, вероятность того, что порог события l окажется к моменту времени t достигнутым или превзойденным, можно определить следующим образом:

$$Q(l,t) = 1 - P(l,t). \quad (6)$$

Анализ показывает, что $\rho_1(x,t)$ и $\rho_2(x,t)$ при любых значениях t и x не являются отрицательными, для функции $Q(l,t)$ при $t \to \infty$ выполняется условие $Q(l,t) \to 1 (\mathbf{P}(l,t) \to 0)$.

5. Экспериментальная проверка предлагаемой модели прогнозирования событий новостной ленты

5.1. Определение параметров модели прогнозирования событий на основе изменения структуры кластеров в информационном пространстве новостных лент

Для моделирования тематического события в новостной ленте на основе разработанной модели необходимо определение ее параметров (ξ и ε). Модель прогнозиро-

вания информационных событий в новостных лентах на основе решения краевой задачи для систем с реализацией памяти и самоорганизацией, основывается на использовании параметров, учитывающих возможность уменьшения текущей величины состояния системы: (тренд уменьшения ξ) и увеличения (тренд увеличения ε). Данные параметры связаны с динамикой изменения структуры новостных кластеров, и могут быть определены на ее основе.

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. 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, создаем вектора и разделяем их на W кластеров по различным темам (в нашем случае W = 300, т. е. оказалось 299 тематических плюс один кластер в который попали все, не вошедшие в 299 тематических кластера, тексты). Далее каждый из W кластеров разделяем 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 ее параметров будем использовать текстовое описание тематического события, произошедшего в какой-либо из дней 2017 года. Берем из 2017 года известное новостное сообщение (опубликованный текст с датой реализации события, описанного в новости). Создаем его вектор N_i.

3) Для каждого дня 2016 года, внутри каждой дневной подгруппы векторов каждого кластера, определяем координаты центроидов:

$$C_j(t) = \left\{c(t)_{1,j}, c(t)_{2,j}, ..., c(t)_{k,j}, ..., c(t)_{M,j}\right\},$$

где $c(t)_{k,j}$ — среднее арифметическое координат входящих в подгруппу векторов (для данного дня без накопления за предыдущие периоды) в момент времени t, а j принимает значения от 1 до W (т. е. получаем для каждого дня получим 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 находим для каждого момента времени $t = \overline{1, 365}$ (для каждого дня) величины косинусов углов между векторами дневных центроидов $C_j(t)$ и вектором новости N_i текстового описания прогнозируемого события (обозначим эти косинусы, как $S_j(t) = \cos\{C_j(t); N_i\}$). Если в данный день новостей нет, то косинусная метрика будет равна пустому множеству.

5) Выбираем отличные от пустого множества значения косинусной метрики и соответствующие им дни 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 ее 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 $[\Delta_j(t_2 - t_1) = (S_j(t_2) - S_j(t_1))/(t_2 - t_1)$ — оно может быть, как положительным, так и отрицательным). Затем берем третье $(S_j(t_3))$ ненулевое 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, находим приведенное к одному дню (τ = 1) отклонение $[\Delta_j(t_3 - t_2) = (S_j(t_a) - S_j(t_2))/(t_a - t_2)$ — оно может быть, как положительным, так и отрицательным]. Проводим эту процедуру по всем ненулевым значениям косинусной метрики до конца года.

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 $\overline{\Delta_j(\Delta t)} < 0$ примем 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) примем за начальное состояние системы x_0.

5.2. Оценка величины косинусной меры порога реализации события в информационном пространстве новостных лент

Для оценки величины косинусной меры порога реализации события рассмотрим 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». Составим табл. 1 нормализованных (лематизированных) слов данных предложений.

Вычисление косинусной метрики дает значение равное 0.61. В данном случае рассмотрены очень короткие тексты, имеющие большое смысловое сходство. По мере увеличения длины текстов значение косинусной метрики будет существенно уменьшаться, хотя их смысловое значение будет оставаться очень близким, поэтому можно принять, что l = 0.5.

Таблица 1. Нормализованные слова предложений тестового примера

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

5.3. Моделирование зависимости от времени вероятности реализации прогнозируемого события. Анализ результатов моделирования

Для проверки модели в качестве прогнозируемого события были случайным образом выбраны 5 новостей (см. прил. 1), описывающие события, произошедшие в 2017 году. Далее, используя алгоритм описанный в п. 5.3 и созданные текстовые кластеры ($W = 300$) для каждого прогнозируемого события новостной ленты были определены значения параметров модели ξ, ε и x_0 (при нахождении ξ и ε было использовано $\tau = 1$ день), см. прил. 1.

Экспериментально рассчитанные параметры модели, представленные в прил. 2, показывают, что во всех рассматриваемых случаях $\varepsilon = \xi$. Это приводит к тому, что и $a_1 = (\varepsilon - \xi)/\tau = 0$ и уравнения (4a) и (4b) переходят в уравнения:

$$\rho_1(x,t) = -\frac{2}{L}\sum_{n=1}^{M}\frac{1}{\cos(\pi n)}\sin\left(\pi n\frac{x_0}{L}\right)\sin\left(\pi n\frac{L-x}{L}\right)e^{-\frac{a\pi^2 n^2}{L^2}t}, \text{ при } x \ge x_0; \qquad (7a)$$

$$\rho_2(x,t) = -\frac{2}{L}\sum_{n=1}^{M}\frac{1}{\cos(\pi n)}\sin\left(\pi n\frac{L-x_0}{L}\right)\sin\left(\pi n\frac{x}{L}\right)e^{-\frac{a\pi^2 n^2}{L^2}t}, \text{ при } x < x_0. \qquad (7b)$$

Если сравнить смысловое содержание в прил. 1 текстов 1 и 3, то можно заметить, что они очень близки (обе новости описывают криминальные происшествия). Очень важно отметить, что экспериментально рассчитанные значения параметров модели ξ, ε и x_0 для них оказываются одинаковыми, что косвенно подтверждает правильность используемой модели.

Результаты моделирования зависимости от времени вероятности реализации прогнозов для событий, описанных в прил. 1 с использованием уравнений (5)–(7) и определенными на множестве из 300 кластеров наборами параметров модели, представлены на рис. 2 (номер кривой соответствует номеру текста в прил. 1). Крупные черные точки на рис. 2 соответствуют моментам времени фактической реализации событий. Полученные результаты показывают, что разработанная модель прогнозирования событий новостной ленты является адекватной и непротиворечивой (все описанные новостные события в зависимости от того, какая глубина памяти учитывается реализовались при высоких значениях вероятности (около 0.8).

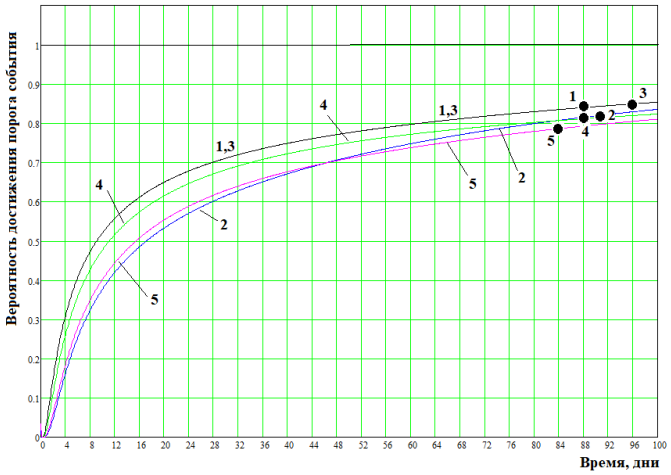

Рисунок 2. Результаты моделирования преодоления порога событий для пяти новостей, описанных в Приложении 1 (*l* = 0.5) *для простой диффузионной модели*

Представляется интересным провести проверку разработанной модели на способность прогнозирования фиктивной новости (то, чего не может происходить на самом деле). В качестве примера можно взять небольшой отрывок из русской народной сказки про Колобка.

«Жили-были старик со старухой. Вот и говорит старик старухе:

— Поди-ка, старуха, по коробу поскреби, по сусеку помети, не наскребешь ли муки на колобок.

Взяла старуха крылышко, по коробу поскребла, по сусеку помела и наскребла муки горсти две.

Замесила муку на сметане, состряпала колобок, изжарила в масле и на окошко студить положила.

Колобок полежал, полежал, взял, да и покатился — с окна на лавку, с лавки на пол, по́ полу к двери, прыг через порог — да в сени, из сеней на крыльцо, с крыльца на двор, со двора за ворота, дальше и дальше.

Катится Колобок по дороге, навстречу ему Заяц:

— Колобок, Колобок, я тебя съем!

— Не ешь меня, Заяц, я тебе песенку спою:

Я Колобок, Колобок,
Я по коробу скребен,
По сусеку метен,
На сметане мешон
Да в масле пряжон,
На окошке стужон.
Я от дедушки ушел,
Я от бабушки ушел,
От тебя, зайца, подавно уйду!

И покатился по дороге — только Заяц его и видел!»

Далее, используя алгоритм описанный в «Определение параметров модели прогнозирования событий на основе изменения структуры кластеров в информационном пространстве новостных лент» и ранее созданные текстовые кластеры 2016 г. ($W = 300$) для данного прогнозируемого события определим значения параметров модели ξ, ε и x_0 (при нахождении ξ и ε было использовано $\tau = 1$ день), см. табл. 2.

Таблица 2 – Нормализованная новость про «колобка» и параметры модели для ее прогнозирования

Нормализованный текст новости	*Дата события*	*Величина параметра* ε	*Величина параметра* ξ	*Начальное состояние системы* x_0 *31.12.2016*
{"id":"85e74845-70da-434c-a602-497efa002de6","date":1514753700000,"title":"Колобок","content":"бабушка ворот говорить горсть два дверь дедушка дорога жить-быть замесить изжарить катиться крылышко метеный мешон навстречу окно песенка подавно поди-ка покатиться пол половина положить помело помести порог прыг пряжон скребена состряпать спеть студить стужон съесть через взять далекий двор крыльцо лавка масло наскрести окошко полежать поскрести сени сметана старик взять далекий двор крыльцо лавка масло наскрести окошко полежать поскрести сени сметана старик заяц короб мука сусек уйти заяц короб мука сусек уйти заяц короб мука сусек уйти старуха старуха старуха старуха колобок колобок колобок колобок колобок колобок колобок колобок","url":"http://null.ru/null","siteType":"Фиктивная"}	Срок реализации не известен	0.0022	0.0022	0.0076

Используя результаты моделирования реализации реальных событий новостной ленты с использованием простой диффузионной модели, для приемлемой вероятности реализации событий можно принять величину равную 0.8 (по сути, эта величина является калибровкой для величины вероятности реализации события, при которой уже следует рассматривать возникновение события). Это позволяет оценить время реализации данного события (при заданной вероятности). Моделирование динамики вероятности реализации новости про Колобка с течением времени для разработанной модели дает оценку времени его реализации (при величине вероятности 0.8 около 90000 дней ≈ 240 лет), что является маловероятным для реализации события новостной ленты. Таким образом, пример с фиктивной новостью также показывает, что разработанная модель прогнозирования событий в новостной ленте является адекватной и непротиворечивой (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).

5.4. Оценка точности и достоверности прогнозов реализации событий в новостной ленте, полученных на основе разработанной модели динамики контента новостных лент

Проблема при оценке точности и достоверности прогнозов реализации событий в новостной ленте на основе разработанной модели заключается в том, что для экспериментальной проверки прогноза имеется только одна наблюдаемая реализация события с известным временем, когда оно уже произошло, и невозможно осуществить множественность испытаний его возникновения.

При прогнозировании значений физически измеряемых величин точность прогноза тем больше, чем меньше величина ошибки, которая представляет собой разность между прогнозируемым и фактическим значениями исследуемой величины. В случае прогнозирования новостных событий, мы имеем дело с зависимостью от времени вероятности того, что описываемое событие может произойти, т. е. в каждый момент времени имеется значение вероятности осуществления (или не осуществления) прогнозируемого события. Экспериментально наблюдаемой физической величиной является только время, когда данное событие осуществляется. Нет параметра, который мог бы измерять величину события (рубли, килограммы, метры и т. д.). Полученные теоретические функции распределения позволяют оценить значение вероятности реализации события, соответствующие данному времени, а определение точности и достоверности прогноза события в новостной ленте по одной наблюдаемой его реализации является неоднозначной задачей в том смысле, что событие может произойти и при очень маленькой величине вероятности и может еще не произойти при вероятности близкой к единице, но возможности провести серию испытаний нет.

Точность прогноза можно оценить величиной доверительного интервала для заданной вероятности его осуществления, а для оценки достоверности необходимо рассчитать вероятности осуществления прогноза в заданном доверительном интервале.

Разработанная модель прогнозирования событий оперирует теоретическими функциями зависимости от времени плотности вероятности его возникновения. Для приближенной оценки точности прогнозирования для пяти новостей, описанных в таблице 3, используем следующий подход. В момент реализации этих событий вероятности их наблюдения становятся равными 1, а расчетные вероятности имеют значения меньше 1. Формально можно сказать, что погрешность определе-

ния вероятности реализации этих новостных событий лежит в диапазоне от 0.22 (22%) до 0.15 (15%), см. рис. 2.

Таким образом, точность будет составлять от 80 до 85%, такова же будет примерно и оценка достоверности.

5. Выводы

Разработанная модель прогнозирования событий в новостной ленте является адекватной и непротиворечивой (все новостные события, использованные для проверки модели, реализуются при высоких значениях вероятностей (около 80%), или если они являются фиктивными, то могут реализоваться только за неприемлемо большое время).

Анализ модели прогнозирования событий в новостной ленте на основе простой диффузионной модели подтверждает возможность прогнозирования событий новостной ленты исходя из их 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 (в том числе при больших временах вероятности асимптотически стремятся к единице).

Литература

[1] *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.

[2] *Gruhl D., Guha R., Kumar R., Novak J., Tomkins A.* The predictive power of online chatter. // KDD ‘05: Proceeding of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. — New York : ACM Press, 2005. P. 78–87.

[3] *Mishne G., Rijke M. D.* Capturing global mood levels using blog posts // AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs. Menlo Park, Stanford University. — The AAAI Press, 2006. P. 145–152.

[4] *Liu Y., Huang X., An A., Yu X.* ARSA: a sentiment-aware model for predicting sales performance using blogs // SIGIR ‘07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. — New York : ACM, 2007. P. 607–614.

[5] *Choi H., Varian H.* Predicting the Present with Google Trends // Tech. rep. — Google, 2009.

[6] *Zhao L., Sun Q., Ye J., Chen F. and et al.* Multi-task learning for spatio-temporal event forecasting // Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD. — New York, 2015. P. 1503–1512.

[7] *Achrekar H., Gandhe A., Lazarus R., Yu S.-H., Liu B.* Predicting flu trends using twitter data. // IEEE Conference on Computer Communications Workshops. — IEEE, 2011. P. 702–707.

[8] *O'Connor B., Balasubramanyan R., Routledge B. R., Smith N. A.* From tweets to polls: Linking text sentiment to public opinion time series // Proceedings of the Fourth International Conference on Weblogs and Social Media. — The AAAI Press, 2010. P. 122–129.

[9] *Tumasjan A., Sprenger T., Sandner P. Welpe I.* Predicting elections with twitter: What 140 characters' reveal about political sentiment // Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. — The AAAI Press, 2010. P. 178–185.

[10] *Ramakrishnan N., Butler P., Muthiah S., and et al.* "Beating the News" with EMBERS: Forecasting civil unrest using open source indicators // Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD. — New York, 2014. P. 1799–1808.

[11] *Ning Y., Muthiah S., Rangwala H., Ramakrishnan N.* Modeling precursors for event forecasting via nested multi-instance learning // Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. — 2016. P. 1095–1104.

[12] *Chouhan S. S., Khatri R.* Data Mining based Technique for Natural Event Prediction and Disaster Management // *International Journal of Computer Applications*. 2016. Vol. 139. No. 4. P. 34–39.

[13] *Орлов Ю. Н., Шагов Д. О.* Индикативные статистики для нестационарных временных рядов // *Препринты ИПМ им. М. В. Келдыша*. 2011. № 53.

[14] *Гнеденко Б. В.* Курс теории вероятностей. — М. : Физматлит, 1961.

[15] *Fuentes M.* Non-Linear Diffusion and Power Law Properties of Heterogeneous Systems: Application to Financial Time Series // *Entropy*. 2018. Vol. 20. No. 9. P. 649.

[16] *Орлов Ю. Н., Федоров С. Л.* Генерация нестационарных траекторий временного ряда на основе уравнения Фоккера // *Планка. Труды МФТИ*. 2016. Т. 8. № 2. С. 126–133.

[17] *Radinsky K., Horvitz E.* Mining the web to predict future events // Proceedings of the 6 th ACM International Conf. on Web Search and Data Mining. — ACM, 2013. P. 255–264.

[18] *Gunawardana A., Meek C., Xu P.* A model for temporal dependencies in event streams // Advances in neural information processing systems. — 2011. P. 1962–1970.

[19] *Manning C. D., Raghavan P., Schütze H.* Introduction to Information Retrieval. — Cambridge University Press, 2008.

[20] *Tan P.-N., Steinbach M., Kumar V.* Introduction to Data Mining. — Pearson Addison-Wesley, 2006.

[21] *Andrews N. O., Fox E. A.* Recent developments in document clustering. — Virginia Polytechnic Institute & State University, 2007.

[22] *Feldman R., Sanger J.* The Text Mining Handbook. — Cambridge University Press, 2009.

[23] *Amdahl G. M.* Validity of the single processor approach to achieving large scale computing capabilities // Proceedings of the April 18–20, 1967, Spring Joint Computer Conference (AFIPS '67 (Spring)) — New York : ACM, 1967. P. 483–485.

Приложение 1. Некоторые нормализованные новостные события 2017 г. и параметры модели для их прогнозирования

№	*Нормализованный текст новости*	*Дата события*	*Величина параметра* ε	*Величина параметра* ξ	*Начальное состояние системы* x_0 *31.12.2016*
1.	**Текст 1**	29.03.2017 (срок реализации 88 дней)	0.016	0.016	0.046
2.	**Текст 2**	29.03.2017 (срок реализации 88 дней)	0.021	0.021	0.083
3.	**Текст 3**	06.04.2017 (срок реализации 96 дней)	0.016	0.016	0.047
4.	**Текст 4**	01.04.2017 (срок реализации 91 день)	0.011	0.011	0.036
5.	**Текст 5**	25.03.2017 (срок реализации 84 дня)	0.016	0.016	0.060

Текст 1:

{"id":"9dc7c737-0359-418f-a809-28a4aa23b3bb","date":1490774096000,"title":"Главу управления МВД убили после выявленных им хищений на 10 млрд","content":"пара неделя покушение николай волк написать заявление увольнение собственный желание отказываться подписывать инвентаризация внутренний финансовый отчет информация лайф убитый накануне глава ресу мвд николай волков жаловаться родной ведомство похищать актив миллиард рубль заставлять подписывать документ чистый показание родные свидетель проверять следователь ск россия непосредственный убийца разыскивать оперативник центральный управление гуур мвд россия источник редакция сообщать волк выявлять многомиллиардный хищение назначать множество внутренний проверка число инвентаризация подозрение подтверждаться устанавливать следствие человек потребовать высокопоставленный полицейский подписание акт проверка ресу мвд якобы никакой финансовый дыра хищение ресу мвд смочь вовремя рассчитываться подрядчик задолжать многий организация ранее мвд возбуждать дело факт мошенничество против организация мариотрек ответственный строительство санаторий министерство олимпиада сочи фгуп ресу мвд выступать заказчик услуга речь идти махинация миллион рубль выявление факт причастность сотрудник ресу мвд махинация дело передавать следственный комитет известно данный момент мвд оставаться должно сочинский строитель минимум миллион рубль ресу мвд являться ответчик арбитражный дело иск ооо строй универсал долг миллион рубль организация ооо предприятие РЦПП ресу задолжать миллион мвд россия комментарий данный ситуация отказываться напоминать убийца преследовать цель ограбить волков забирать портфель деньги оставлять место дорогостоящий телефон наличный деньги киллер скрываться автомобиль ваз забывать место медицинский маска ск рассматривать заказной убийство приоритетный версия гибель глава ресу мвд","url":" https://life.ru/991216 ","siteType":"LIFE"}

Текст 2:

{"id":"3845f74e-c144-4ec3-9b8f-333e8e08b8ad","date":1490776169000,"title":"Таджикистан стал главным зарубежным поставщиком смертников для ИГИЛ","content":"вывод приходить автор исследование война посредством самоубийство статистический анализ индустрия мученичество исламский государство иго опубликовывать международный центр борьба терроризм гааг нидерланды период декабрь год ноябрь год лишь смертник иго управлять нагружать взрывчатка машина ингимаси боец пояс смертник воевать обычный оружие необходимость подрываться рядом враг прима лайф живой бомба дом указывать иностранный боец отмечать автор исследование общий

сложность иностранец погибать качество смертник рассматривать год пятнадцать упоминать куний принимать исламский традиция прозвище связанный место происхождение прима лайф аль мухаджир подобно аль ансари указывать иностранец указывать страна происхождение оставаться погибать качество управлять машина взрывчатка происходить страна таджикистан затем идти выходец саудовский аравия марокко тунис россия далее приводить таблица указывать точный цифра смертник иго таджикистан саудовский аравия марокко тунис россия странно год многочисленный иммигрировать салафит тунис являться крупный иностранный легион иго насчитывать около тыс боец вплотную идти тыс выходец ваххабитский королевство ас сауд уроженец основывать вслед следовать иммигрант иордания править королевский династия принадлежать род хашимит происходить прадед пророк мухаммед возможно поэтому список смертник указывать период лишь иорданец марокканец двенадцать месяц идти речь значительно таджик погибать сирия ирак ход атака нагружать взрывчатка машина ингимаси выходец зарубежный страна отмечать представитель международный центр борьба терроризм цифра поразительный рассматривать душа население количество выходец различный страна ряд иго прима лайф предполагать таджик часто направляться самоубийственный подрыв минимум частично национальность организация запрещать россия верховный суд рф","url":" https://life.ru/991022 ","siteType":"LIFE"}

Текст 3:

{"id":"5fbf3918-22cc-4ef3-8ad0-20ae2654286c","date":1491441192000,"title":"В районе нападения на сотрудников Росгвардии в Астрахани идет боестолкновение","content":"сообщать лайф источник правоохранительный орган ленинский район астрахань начинаться боестолкновение злоумышленник предположительно несколько час ранее нападать сотрудник росгвардия предварительный данные спецоперация проходить район железнодорожный станция астрахань уточнять источник напоминать сегодня ночь трое росгвардеец получать огнестрельный ранение нападение несколько преступник заявлять региональный управление ск рф атака боец росгвардия причастный злоумышленник апрель убивать сотрудник полиция астрахань","url":" https://life.ru/994664 ","siteType":"LIFE"}

Текст 4:

{"id":"c7584973-348d-417a-90c3-2199a4040558","date":1491047117000,"title":"НАТО не собирается воевать с Россией из-за Абхазии и Южной Осетии","content":"представитель нато южный кавказ уильям лахью заявлять блок воевать россия абхазия южный осетия случай вступление грузия североатлантический альянс грузия должный решать статус территория четко понимать пока стоять российский войско пятый статья грузия использовать никто хотеть война лахью сообщать член альянс договариваться грузия член нато никакой срок возможный вступление грузия альянс называть сообщать интерфакс потихоньку дело идти вперед будущее грузия получать приглашение знать лахью слово вступление грузия нато зависеть параллельный фактор политика разный страна готовность грузия","url":" http://www.vesti.ru/doc.html?id=2872818 ","siteType":"VESTI"}

Текст 5:

{"id":"dacb1299-f6fa-4b25-a4cd-95795657cf4c","date":1490474466000,"title":"Сирийские военные с января освободили от ИГ* 195 населенных пунктов","content":"число населенный пункт освобождать январь сирийский правительственный войско террористический организация исламский государство иго январь достигать сообщать суббота российский центр примирение враждовать сторона сирия количество населенный пункт освобождать январь год сирийский правительственный войско вооруженный формирование международный террористический организация исламский государство увеличиваться говориться бюллетень опубликовывать сайт минобороны рф сутки контроль правительственный войско переходить квадратный километр территория общий сложность освобождать квадратный километр количество населенный пункт присоединяться процесс примирение сутки изменяться сообщение центр примирение продолжаться переговоры присоединение режим прекращение боевой действие отряд вооруженный оппозиция провинция алеппо дамаск хам хомс эль кунейтр число вооруженный формирование заявлять прекращение боевой действие соответствие соглашение перемирие изменяться террористический организация запрещать россия","url":" https://ria.ru/syria/20170325/1490808936.html ","siteType":"RIA"}

Автор:

Ольга Александровна Новикова — ассистент кафедры информационных технологий в государственном управлении, МИРЭА — Российский технологический университет

Modeling and Forecasting the Dynamics of Events in News Feeds Based on a Simple Diffusion Model

O. A. Novikova

MIREA – Russian Technological University, 78 Vernadsky Avenue, Moscow 119454

e-mail: novikova@mirea.ru

Abstract. One of the problems of predicting events in news feeds is the development of models that allow working with a weakly structured information space of text documents. The article describes a model for predicting events in news feeds based on stochastic dynamics of changes in the structure of non-stationary time series of news clusters (States of the information space) based on the application of the diffusion approximation. The forecasting of news feed events are carried out on their textual description, vectorization, and finding the value of the cosine of the angle between this vector and the centroids of various semantic clusters of the information space. The change in the value of this cosine over time can be examined as the wandering of a point on the segment [0,1], which contains a trap at the point of the event realization threshold, where the wandering point can get over time. In creating the model, the probability schemes of transitions between states in the information space were considered. Based on this approach, a nonlinear second-order differential equation was derived, and a boundary value problem for predicting news events was formulated and solved. All this allowed us to get a theoretical time dependence of the probability density function for the distribution of parameters of non-stationary time series describing the evolution of the information space. The results of probability modeling of event realization depending on time (with experimentally defined sets of parameters of the developed model for already implemented events) show that the model is consistent and adequate (all news events used to test the model, depending on what memory depth is taken into account, are realized at high probability values (approximately 80%), or if they are fictitious, they can be realized only for an unacceptably long time).

Keywords: forecasting events in news feeds, news cluster, news clustering, stochastic dynamics of changes in information system states, threshold for news events.

References

[1] *Nivikova O. A.*, *Andrianova E. G.* (2018) *Cloud of Science*, **5**(1):196–211. [Rus]

[2] *Gruhl D.*, *Guha R.*, *Kumar R.*, *Novak J.*, *Tomkins A.* (2005) The predictive power of online chatter. In KDD '05: Proceeding of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (ACM Press), pp. 78–87.

[3] *Mishne G.*, *Rijke M. D.* (2006) Capturing global mood levels using blog posts. In AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (Stanford University), pp. 145–152.

[4] *Liu Y.*, *Huang X.*, *An A.*, *Yu X.* (2007) ARSA: a sentiment-aware model for predicting sales performance using blogs. In SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM), pp. 607–614.

[5] *Choi H.*, *Varian H.* (2009) *Predicting the Present with Google Trends* (Tech. rep.; Google).

[6] *Zhao L.*, *Sun Q.*, *... & Ramakrishnan N.* (2015) Multi-task learning for spatio-temporal event forecasting. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD, pp. 1503–1512.

[7] *Achrekar H., Gandhe A., Lazarus R., Yu S.-H., Liu B.* (2011) Predicting flu trends using twitter data. In IEEE Conference on Computer Communications Workshops, pp. 702–707.

[8] *O'Connor B., Balasubramanyan R., Routledge B. R., Smith N. A.* (2010) From tweets to polls: Linking text sentiment to public opinion time series. In Proceedings of the Fourth International Conference on Weblogs and Social Media, pp. 122–129.

[9] *Tumasjan A., Sprenger T., Sandner P. Welpe I.* (2010) Predicting elections with twitter: What 140 characters' reveal about political sentiment. In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, pp. 178–185.

[10] *Ramakrishnan N., Butler P., ... & Kuhlman C.* (2014) "Beating the News" with EMBERS: Forecasting civil unrest using open source indicators. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD, pp. 1799–1808.

[11] *Ning Y., Muthiah S., Rangwala H., Ramakrishnan N.* (2016) Modeling precursors for event forecasting via nested multi-instance learning. In // *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. —. P. 1095–1104.

[12] *Chouhan S. S., Khatri R.* (2016) *International Journal of Computer Applications*, **139**:34–39.

[13] *Orlov Y. N., Shagov D. O.* (2011) *Indicative statistics for non-stationary time series* (Keldysh Institute preprints, Vol. 53). [Rus]

[14] *Gnedenko B. V.* (1961) *Kurs teorii veroyatnostej* (Fizmatlit). [Rus]

[15] *Fuentes M.* (2018) *Entropy*, **20**(9):649.

[16] Orlov Y. N., Fedorov S. L. (2016) *Planka. TRUDY MFTI*, **8**(2):126–133. [Rus]

[17] *Radinsky K., Horvitz E.* (2013) Mining the web to predict future events. In Proceedings of the 6 th ACM International Conf. on Web Search and Data Mining, pp. 255–264.

[18] *Gunawardana A., Meek C., Xu P.* (2011) A model for temporal dependencies in event streams. In Advances in neural information processing systems, pp. 1962–1970.

[19] *Manning C. D., Raghavan P., Schütze H.* (2008) *Introduction to Information Retrieval* (Cambridge University Press).

[20] *Tan P.-N., Steinbach M., Kumar V.* (2006) *Introduction to Data Mining* (Pearson Addison-Wesley).

[21] *Andrews N. O., Fox E. A.* (2007) *Recent developments in document clustering* (Virginia Polytechnic Institute & State University).

[22] *Feldman R., Sanger J.* (2009) *The Text Mining Handbook* (Cambridge University Press).

[23] *Amdahl G. M.* (1967) Validity of the single processor approach to achieving large scale computing capabilities. In Proceedings of the April 18–20, 1967, spring joint computer conference (AFIPS '67 (Spring)), pp. 483–485.