УДК 519.21

Сигов А.С., Жуков Д.О., Новикова О.А.

Московский технологический университет, г. Москва, Россия

МОДЕЛИРОВАНИЕ ПРОЦЕССОВ РЕАЛИЗАЦИИ ПАМЯТИ И САМООРГАНИЗАЦИИ ИНФОРМАЦИИ ПРИ ПРОГНОЗИРОВАНИИ НОВОСТНЫХ СОБЫТИЙ С ИСПОЛЬЗОВАНИЕМ МАССИВОВ ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ТЕКСТОВ

АННОТАЦИЯ

В работе представлена разработанная модель прогнозирования новостных событий на основе стохастической динамики изменения кластеров новостных образов и реализации памяти в информационном пространстве при самоорганизации слабоструктурированной информации. Проведённый авторами работы анализ стохастической динамики достижения порога реализации новостного события показывает возможность роста вероятности перехода через него практически сразу после начала процесса изменения структуры новостных кластеров, что связано с учетом памяти о предыдущих состояниях в информационной системе и возможности описания самоорганизации вследствие учета в дифференциальной модели информационных процессов второй производной по времени. Кроме того, предлагаемая модель показывает возможность резких изменений вероятности перехода через порог событий, и учитывает наличие в её поведении осцилляций. На основе разработанной модели создан алгоритм анализа взаимосвязи новостных кластеров в информационном пространстве с возможностью возникновения прогнозируемого события, и определения возможного времени его реализации.

КЛЮЧЕВЫЕ СЛОВА

Самоорганизация; случайные процессы с памятью; порог редкого события; информационное пространство; новостной кластер.

Sigov A.S., Zhukov D.O., Novikova O.A.

Moscow technological university, Moscow, Russia

MODELLING OF MEMORY REALIZATION PROCESSES AND THE IMPLEMENTATION OF INFORMATION SELF-ORGANIZATION IN FORECASTING THE NEW'S EVENTS USING ARRAYS OF NATURAL LANGUAGE TEXTS

ANNOTATION

The paper presents a developed model of forecasting of news events on the basis of the stochastic dynamics of the news clusters and realization of memory in the information space with the self-organization of semistructured information. The stochastic dynamics of achieving the threshold of realization of the news event analysis, carried out by the authors, shows the possibility of growth of probabilities of transition through the realization threshold probabilities almost immediately after the beginning of the process of modification in the structure of the news cluster, the probability of transition is determined by previous states in the information space memory consideration and the possibility of self-organization description, resulting from information processes, the second time derivative, in the differential equation consideration. In addition, the proposed model shows the possibility of sudden changes in the probability of passing the events threshold, and takes into account the presence of oscillation in its behavior. On the basis of the developed model the algorithm of analysis of connection between news clusters in the information field and the possibility of occurrence of the predicted event as well as determining the possible time of its implementation has been created.

KEYWORDS

Self-organization; stohastic processes with memory; the threshold is a rare event; information space; a news cluster.

Быстрый рост объёмов баз данных во всех сферах человеческой деятельности привел к чрезвычайной востребованности эффективных инструментов по преобразованию данных в ценную информацию. Обширные объемы текстовой информации, доступной в электронном виде, продолжают расти с ошеломляющей скоростью. По оценкам экспертов к 2020 году общемировой объем различных данных будет около 35000 Экзобайт (Экзо=10^{18}), что составит рост по отношению к 2010 году в 44 раза. Причем до 90% будут составлять не структурированные или плохо структурированные данные, а до 80% данных не будут использованы повторно.

Представленная работа посвящена разработке принципов и научных методов прогнозирования редких антропогенных событий, способных оказывать существенное влияние на социальное и экономическое развитие общества.

В настоящее время существует выдвинутая Нассимом Николасом Талебом теория, в которой рассматривается природа возникновения и реализации таких событий. Согласно его теории, данные явления должны удовлетворять следующим критериям [1]:

1. Событие является неожиданным (для экспертов);
2. Событие имеет значительные последствия;
3. После наступления, в ретроспективе, событию можно дать рационалистическое объяснение, как если бы оно было ожидаемым.

Нассим Николас Талеб для таких явлений ввел понятие: «Чёрный лебедь» («Black swan»). С его точки зрения практически все значимые научные открытия, исторические и политические события, достижения искусства и культуры — это «Чёрные лебеди». В своей теории Нассим Талеб предполагает, что человечество не способно успешно прогнозировать своё будущее, а уверенность в своих знаниях опережает сами знания и порождает феномен «сверхуверенности» [1].

Поскольку в реальном мире существуют причинно–следственные связи, то на наш взгляд нельзя однозначно утверждать, что данная задача не имеет решения, и в представленной нами работе мы обозначаем некоторые возможные пути её решения. Создание моделей возникновения «Чёрных лебедей», позволяющих предсказывать возможность подобного рода событий, представляет на наш взгляд значительный как научный, так и практический интерес.

Частичное решение задачи прогнозирования событий достигнуто в экономических областях, например, для технического анализа поведения фьючерсных рынков [2]. Для этого в частности можно использовать многомерный статистический анализ и временные ряды [3]. Обсуждая тему математического моделирования и прогнозирования развития экономических процессов необходимо обратить внимание на работы российской научной школы академика А.А. Петрова и чл. – корр. И.Г. Поспелова [4-6], работы профессора Ф.Т. Алескерова [7,8], посвященные теории выбора (модель интервального выбора, модель коллективного выбора, основанные на рассмотрении классов бинарных отношений, таких как слабые бипорядки, простые и простейшие полупорядки), на основе которой может быть построена поведенческая модель биржи.

В настоящее время основные работы в области прогнозирования социальных информационных процессов направлены на описание поведения пользователей в социальных сетях и сети интернет. В частности, можно упомянуть работы [9,10].

Модели, близкие по проблематике к предлагаемой нами теме прогнозирования новостных событий, были рассмотрены в работах [11,12]. В [11] для решения задачи прогнозирования 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. Однако, авторы не рассматривают наличие памяти информационного пространства и возможность его самоорганизации, что является на наш взгляд очень важным для прогнозирования редких событий. В работе [12] для описания структуры новостного информационного пространства вводится понятие директора - условной оси, положение которой определяется усреднением направлений векторов, задающих положение центров всех новостных кластеров. Авторы предлагают проводить анализ самоподобия в поведении директоров на основе модели Хёрста и на основании этого выявлять наличие периодичности в их поведении, и прогнозировать повторяемость.

Отмечая существенные успехи, достигнутые в прогнозировании поведения процессов для технических и экономических систем, нельзя не обратить внимание на то, что для социальных систем существенный прогресс в данной области отсутствует, а механистическое перенесение методов и подходов технических и экономических наук не дает желаемых результатов.

Прогнозирование появления в социальных системах новостных событий типа «Чёрный лебедь» осложняется необходимостью поиска скрытых закономерностей в больших объемах слабоструктурированной гетерогенной информации и решением проблемы измеримости

параметров различных протекающих процессов. Априори все данные являются по своему характеру гетерогенными (имеют разный формат представления и единицы измерения). Важным является то, что использование существующих методов и подходов интеллектуального анализа данных (Big Data) не приносит ощутимых результатов, поскольку применение имеющихся инструментов очистки, извлечения и анализа данных, скорее всего не позволяют извлекать сами знания, уничтожая их скрытые (латентные) связи (для изучения нелинейных процессов применяются линейные методы).

Появление редкого новостного события типа «Чёрный лебедь» может зависеть от факторов любой природы (природных, климатических, астрономических, геологических, социальных и т.д.). Например, установлено влияние солнечной активности (11 летний и 60 летний циклы) на социальные (войны, революции, волны переселения народов и т.д.) и экономические процессы (неурожай, голод, падение биржевых индексов и т.д.). Климатические условия и их изменения также влияют в исторической перспективе на социально – экономические процессы.

При создании модели прогнозирования новостных событий необходим математический аппарат, который позволил бы формализовать характер данных и привести их к единой шкале измерений. Очевидно, что нельзя в одной модели проводить вычислительные операции, например, над лингвистическими оценками и величинами метрической шкалы, без использования процедур отображения на формальное безразмерное множество.

Суть предлагаемых нами принципов для возможности прогнозирования новостных событий заключается в следующем:

1. На основе существующих методов математической лингвистики необходимо формализовать описание информационного пространства, представляющего набор текстов на естественных языках. Таким образом, предполагается решить проблему разнородности данных и единиц измерения параметров различных процессов (шкалы измерений заведомо огрубляются и становятся лингвистическими, но при этом все данные формализуются единым образом).
2. На естественном языке можно дать описание интересующего (предполагаемого) новостного события, для которого будет проводиться прогнозирование его реализации. И на основе методов математической лингвистики осуществляется его формализованное описание в информационном пространстве.
3. Учитывая, что информационное пространство является отображением реального мира, в котором существуют причинно–следственные связи между событиями, то можно предположить выполнение закона их сохранения в информационном пространстве. А в качестве гипотезы для проверки можно выдвинуть идею о том, что существующие в информационном пространстве формализованные текстовые знания могут формировать образ интересующего прогнозируемого события. Если данная гипотеза верна, то может быть построена математическая модель прогнозирования новостного события, описывающая временные зависимости вероятности его реализации.

Предлагаемая нами методика разработки модели прогнозирования новостных событий, включает следующие этапы. Сначала необходимо выделить основные характеристики новостных событий, которые на наш взгляд необходимо учесть при построении математической модели их прогнозирования. Затем выбрать методы математической лингвистики (например, представление документа векторной семантической моделью), позволяющие формализовать описание процессов реального мира с использованием текстов на естественных языках для создания их информационных образов, а для анализа возможности прогнозирования новостного события необходимо разработать модель, учитывающую стохастическую самоорганизацию слабоструктурированной информации и реализацию памяти в массивах естественно-языковых текстовых данных. Результаты её анализа, на наш взгляд, позволят предложить алгоритм прогнозирования, который можно использовать совместно с методами математической лингвистики, как инструмента для проверки разработанной нами прогностической модели.

Основные характеристики новостных событий и предположения, принятые для создания модели прогнозирования

Для создания модели прогнозирования новостных событий нами были выделены следующие их основные характеристики и принят ряд предположений:

1. В характере, времени и месте реализации событий имеется изначальная *неопределенность и нечеткость;*
2. Реализующиеся событие является проявлением *стохастических процессов с изначально неизвестными* законами распределения и их характеристиками (математическое ожидание, дисперсия и т.д.);

3. Событие является следствием того, что в системе, где оно наблюдается, *имеется хотя бы "частичная" память* о некоторых ранее происходящих процессах и их состояниях. В реальном мире существуют множественные причинно-следственные связи, хотя бы частично сохраняющиеся при отображении событий в информационное пространство более чем на одном шаге (правило сохранения причинно-следственных связей при любых отображениях);
4. Событие является следствием того, что система, в которой оно происходит, обладает возможностью к хотя бы частичной *самоорганизации*;
5. Процессы, приводящие к появлению события могут *иметь характер самоподобия*. У каждого из таких событий, как правило, наблюдаются слабо выраженные предшествующие ему аналоги, на основе которых, в ретроспективе, дается рационалистическое объяснение произошедшему;
6. Исходные данные для использования в прогностической модели требуют формализации и приведения к единому формату, поскольку являются не измеримыми (нет методик оценки в единой шкале) и гетерогенными (социально – экономические параметры, геологические, климатические, астрономические). Инструментом реализации данного процесса и проверки адекватности модели могут быть существующие методы математической лингвистики. Например, на основе методик, применяемых при отнесении текстовой информации к определенной группе новостных сообщений;
7. Любое прогнозируемое событие может быть описано в информационном пространстве некоторым текстовым документом, который можно отнести к определенной смысловой группе (кластеру), имеющей свои собственные характеристики (*правило кластеризации информации*). В любой момент времени существует множество различных информационных кластеров (описывающих различные природные, социальные и экономические процессы, явления природы, научные знания, культурные, политические, спортивные, военные и иные новостные события) отображающих основные свойства физического мира и взаимосвязи событий. С течением времени новостные кластеры могут изменяться или исчезать, и эти изменения могут быть описаны в рамках динамических моделей;
8. Прогнозирование новостного события должно быть основано на модели, в которой исходная текстовая информация на естественном языке после формализации с помощью методов математической лингвистики (например, представлении текстовых документов векторной семантической моделью) превращается в безразмерные числовые данные, с которыми можно производить математические операции;
9. Несмотря на то, что прогнозируемое событие является ещё неизвестным, можно искусственно описать его в информационном пространстве, создавая некоторый новостной образ, а затем построить динамическую модель возможной трансформации уже существующих текстовых образов к заданному образу события.

Описание методов формализации текстов естественных языков для использования в предлагаемой модели

Всю совокупность естественно – языковых текстов, описывающих различные природные, социальные и экономические процессы, явления природы, научные знания, культурные, политические, спортивные, военные и иные новостные события и т.д., можно определить, как коллекцию документов.

Для формализованного представления текстовых документов необходимо создание словаря коллекции при проведение лингвистической обработки, которая включает [13, 14]:

- Разметка текста (линеаризация). Необходима для распознавания и извлечения слов – терминов (термов) и составных терминов (термов) из исходного текста. При разметке учитывается пунктуация и переносы слов, и кроме того игнорируются стоп – слова (малоинформативные, но часто встречающиеся в текстах слова, например, предлоги, союзы, причастия, междометия и т.д.);
- Нормализация текста. При нормализации проводится удаление из исходного текста грамматической информации (падежи, числа, глагольные виды и времена, залоги причастий, род и так далее). Слова в разных регистрах, а также их аббревиатуры приводятся к одной форме. Нормализация требует морфологического анализа, распознающего части речи с учетом контекста и многочисленных правил согласования (без него нормализация будет давать значительное количество ошибочных результатов). В ряде случаев нормализация может быть заменена стеммингом или лемматизацией;
- Комментирование текста. Комментирование в частности необходимо потому что, одно и то же слово в тексте может быть употреблено в различных значениях. Комментирования заключаются в добавлении в документ метаданных с информацией о части речи, о наличии у слова других

значений и синтаксический разбор предложения. Синтаксическое комментирование включает сегментацию и разметку частей речи. Это используется для устранения неоднозначности в сокращениях и поиск ассоциаций среди ключевых слов. Комментирование полезно для измерения семантической схожести слов и понятий (для моделей, основанных на матрице слово-контекст) [15].

После разметки, нормализации и комментировании текстов каждый документ коллекции представляется набором термов (терминов), которые в него входят, и их частотами вхождения, что позволяет создать векторную модель документа. При этом порядок следования слов, соответствующих термам, в представлении документа не учитывается. Термы коллекции (словарь терминов) могут сортироваться в соответствии с некими критериями (например, по алфавиту слова, представляющего терм, или на основе критерия морфологической эквивалентности). Векторная модель основывается на гипотезе (**statistical semantics hypothesis**) о том, что: статистические зависимости употребления слов и терминов (термов) могут быть использованы для нахождения заложенного в текст смысла [13, 14]. Согласно гипотезе о скрытых связях, пары слов, которые встречаются в похожих моделях, стремятся иметь близкую семантическую зависимость [16].

Суть векторной модели документа (модель **"bag of terms"**) заключается в следующем [13, 14]. Пусть общее количество имеющихся различных уникальных термов равно m (термы могут быть пронумерованы от 1 до m), а число имеющихся в коллекции текстовых документов равно n (документы могут быть пронумерованы от 1 до n). Пусть имеется документ (обозначим его $\mathbf{X_i}$, i – принимает значения от 1 до n), который содержит некоторое число термов (терминов) (их порядок в наборе не важен). Тогда данному документу можно поставить в соответствие вектор $\mathbf{X_i}=\{\omega_1, \omega_2, \omega_3, \cdots \omega_j, \cdots \omega_m\}$, где первый элемент соответствует числу (частоте) вхождений в документ первого термина, второй –второго, третий – третьего и т.д. Каждая ω_j может принимать любое положительное значение от 0. Все $\mathbf{X_i}$ образуют матрицу, которую называют для всей коллекции: термин – документ (столбцами матрицы являются вектора $\mathbf{X_i}$, а строки определяют частоты вхождения отдельных терминов в документы коллекции). Как правило, большинство значений элементов этой матрицы равны 0 (матрица является разреженной). Это связано с тем, что документы содержит лишь малую долю терминов из всего словаря.

Главная идея векторной модели семантики (vector space model, VSM) – это представление каждого документа коллекции в качестве точки в многомерном пространстве (вектора в векторном пространстве, размерности $\mathbf{R^m}$). Близко лежащие друг к другу точки соответствуют семантически схожим документам.

Векторная модель документа получили свое развитие в основном для решения задач информационного поиска: поиск документа по запросу (и определение релевантности), классификация документов, кластеризация документов.

Информационный поиск с использованием матрицы термин-документ (term–document matrix) основывается на следующей гипотезе: оценивание релевантности документа запросу можно производить путем представления документа и запроса в виде набора терминов (bag of terms).

Гипотеза, использованная в нашей модели, заключается в том, что прогнозируемое событие тоже можно рассматривать в определенной степени как поисковый запрос, в котором его предполагаемое текстовое описание может быть формализовано в виде вектора в векторном пространстве, имеющего размерность $\mathbf{R^m}$. Однако, в отличии от задачи информационного поиска и сопоставления вектора запроса векторам из пространства $\mathbf{R^m}$, в нашей модели мы будем наоборот искать зависимости, описывающие эволюцию векторов из $\mathbf{R^m}$ и определение временной зависимости вероятности формирования на её основе вектора прогнозируемого события.

Прежде чем перейти к детальному описанию предлагаемой нами модели необходимо сказать о том, что представляется целесообразным выделить во всей коллекции документов отдельные смысловые группы текстов (провести тематическое разделение, или кластеризацию текстов). При этом каждая такая группа (кластер) текстовых документов может быть описана вектором, задающим положение центра. Данный подход используется, например, в задачах классификации и отнесение сообщений новостной ленты к определенной группе и весьма эффективен с точки зрения экономии вычислительных ресурсов.

Модель стохастической динамики новостных образов при прогнозировании редких событий в информационном пространстве

Взаимосвязь образов новостных кластеров в информационном пространстве с появлением прогнозируемого события

Возьмём коллекцию текстовых документов. Используя методы математической лингвистики, создадим их векторное представление в информационном пространстве с размерностью $\mathbf{R^m}$. Проведем в какой-то момент времени t их кластеризацию по смысловым группам. Определим вектора (z_1, z_2, z_3, z_k, z_j,), задающие положение центров этих кластеров в данный момент времени. Далее проведем текстовое описание прогнозируемого новостного события и зададим его вектор $\mathbf{X_{bs}}$ в информационном пространстве с размерностью $\mathbf{R^m}$.

Поскольку мы предполагаем, что в информационном пространстве уже имеются некоторые данные о предстоящем новостном событии, то должна существовать, и эволюция имеющихся групп новостных событий, на событие которое мы пытаемся прогнозировать. При описании эволюции на наш взгляд является наиболее приемлемым использование параметров, применяемых в информационном поиске при определении релевантности запросов: нахождение расстояния между векторами и определения косинуса угла между ними.

Выберем в качестве эволюционного параметра - нахождение проекций x_j векторов, задающих положение центров информационных кластеров в данный момент времени z_1, z_2, z_3, z_k, z_j, на направление вектора $\mathbf{X_{bs}}$, определяющего появление прогнозируемого события. Каждая из проекций x_k определяется как произведение величины соответствующего вектора z_k и косинуса угла между направлениями векторов z_k и $\mathbf{X_{bs}}$ ($x_k=z_k*\cos(\alpha_k)$, т.е. по сути дела мы используем косинусную меру, принятую в информационном поиске.

Спустя некоторый интервал времени (назовем его интервал измерения τ_0) величины векторов, задающие положения центров информационных кластеров изменяются на какие-то случайные значения Δ_j (j – обозначает рассматриваемый вектор). Для пояснения, например, вектора z_{01} и z_{02} определяют положение центров новостных кластеров в информационном пространстве в момент времени t, а вектора z_1 и z_2 спустя интервал времени τ_0 (момент времени $t+\tau_0$). В данных случаях $\Delta_1=z_1-z_{01}$ и $\Delta_2=z_2-z_{02}$. Аналогичным образом определяются изменения положения центров для всех кластеров в информационном пространстве за время шага τ_0.

Величины x_{01}, x_{02}, x_1 и x_2 будут задавать значения соответствующих проекций векторов, определяющих положение центров новостных кластеров 1 и 2, на направление вектора прогнозируемого события, в моменты времени t и $t+\tau_0$. Следует отметить, что некоторые величины проекций x_k могут оказаться больше предыдущих значений x_{0k} (для одной и той же группы новостных данных), а некоторые меньше, т.е. в информационном пространстве одновременно сосуществуют два тренда поведения. Один на увеличение значений проекций, другой на их уменьшение, что необходимо учесть в рамках разрабатываемой модели. Введем для любого момента времени понятие среднего значения $\overline{x_i}$ всех величин проекции векторов, задающих положение центров новостных кластеров в информационном пространстве на направление оси прогнозируемого события. Для группы из K новостных кластеров в момент времени t среднее значение $\overline{x_t}$ определяется следующим образом: $\overline{x_t}=\frac{\sum_{j=1}^{K} x_{t,j}}{K}$, где $x_{t,j}$ обозначают соответствующие значения проекций векторов, задающих положение центров новостных кластеров, на направление оси прогнозируемого события, в данный момент времени t. Спустя интервал времени τ_0: $\overline{x_{(t+\tau_0)}}=\frac{\sum_{j=1}^{K} x_{(t+\tau_0),j}}{K}$. Чтобы учесть тренды увеличения и уменьшения значений величин проекций векторов, задающих положение центров новостных кластеров, на направление оси прогнозируемого события можно поступить следующим образом. На основании анализа значений $x_{t,j}$ и $x_{t+\tau,j}$ разделим множество $x_{t,j}$ на две подгруппы, в одной $(x_{t,j})_I$ будут все новостные кластеры, для которых за интервал времени τ_0 произошло уменьшение значений величин проекций $x_{t+\tau,j}$ (обозначим число таких кластеров как R), а во второй $(x_{t,j})_{II}$ – увеличение (обозначим число таких кластеров как K-R), и найдем для каждой из них средние значения ($\overline{(x_{t,j})}_I=\frac{\sum_{j=1}^{R}(x_{t,j})_I}{R}$ и $\overline{(x_{t,j})}_{II}=\frac{\sum_{j=1}^{K-R}(x_{t,j})_{II}}{K-R}$) проекций векторов, задающих положение центров этих новостных кластеров. Далее, мы предлагаем следующий подход к учету трендов увеличения и уменьшения значений величин проекций векторов, задающих положение центров новостных кластеров, на направление оси прогнозируемого события. Поскольку при учете трендов имеет смысл говорить об усредненных величинах, то будем рассматривать переход за интервал времени τ_0 в точку $\overline{x_{(t+\tau_0)}}$ из точки $\overline{(x_{t,j})}_I$, которая находится на оси прогнозирования события правее точки $\overline{x_{(t+\tau_0)}}$и точки $\overline{(x_{t,j})}_{II}$ которая находится левее$\overline{x_{(t+\tau_0)}}$. Сами по себе переходы являются случайными событиями, а их величины можно определить следующим образом: $\xi_t=\overline{(x_{t,j})}_I-\overline{x_{(t+\tau_0)}}$ и $\varepsilon_t=\overline{x_{(t+\tau_0)}}-\overline{(x_{t,j})}_{II}$. После следующего шага τ_0 определяем новые значения $\xi_{t+\tau}$ и $\varepsilon_{t+\tau}$: $\xi_{t+\tau_0}=\overline{(x_{t+\tau_0,j})}_I-\overline{x_{(t+2\tau_0)}}$ и $\varepsilon_{t+\tau_0}=\overline{x_{(t+2\tau_0)}}-\overline{(x_{t+\tau_0,j})}_{II}$ и т.д.

На любом шаге **n** величины $\xi_{t+n\tau}$ и $\varepsilon_{t+n\tau}$ могут иметь различные случайные (или почти случайные) значения. Поэтому возникает необходимость либо выявления в их поведении характерных особенностей (например, зависимости $\xi_{t+k\tau}$ и $\varepsilon_{t+k\tau}$ от времени, могут обладать *самоподобием при не случайном поведении*), либо, если они обладают характеристиками равномерного распределения, то возможно их усреднение за достаточно большой интервал времени наблюдения и использование в модели этих средних случайных значений $\xi = \frac{\sum_{k=1}^{N} \xi_{t+k\tau}}{N}$ и $\varepsilon = \frac{\sum_{k=1}^{N} \varepsilon_{t+k\tau}}{N}$, где N – число шагов (интервалов времени τ_0) наблюдения).Отметим, что зависимости величин $\xi_{t+k\tau}$ и $\varepsilon_{t+k\tau}$ от времени могут подчиняться некоторому закону распределения и тогда его параметры могут быть установлены по полученным данным.

Самоподобие характеристик процессов в информационном пространстве. Медленно и быстро меняющиеся подсистемы

При наличии во временной зависимости поведении величин $\xi_{t+k\tau}$ и $\varepsilon_{t+k\tau}$ элементов самоподобия, для его описания может быть применена теория, суть которой заключается в том, что непрерывный стохастический процесс Y(t), считается статистически самоподобным с параметром Харста (Hurst) H (0,5≤H≤1), если для любого положительного числа **a**, случайные процессы Y(t) и $\mathbf{a}^{-H}$Y(at) будут иметь одинаковые статистические свойства. Значение H=0.5 показывает отсутствие самоподобности, а значения H близкие к единице показывают её большую степень.

Следует отметить, что теория самоподобия широко применяется для исследования информационных процессов, в частности поведении трафика при передаче данных [17].

Самоподобные процессы принято разделять на два класса: точно самоподобные и асимптотически самоподобные. Различие между этими двумя классами заключается в том, что для функции автокорреляции точно самоподобных процессов выполняется условие: $R(Y_k) = R(Y_q)$, а для асимптотически самоподобных: $R(Y_k) \rightarrow R(Y_q)$, при k→q. Дисперсия для обоих классов процессов определяется одинаковым образом: $D(Y_k) = {D(Y_q)}/{k^{\beta}}$, где β – параметр самоподобия (0<β<1), связанный с параметром Харста следующим соотношением: β=2(1–H), а k – величина блока разбиения исходных данных.

Коэффициент Харста находится по зависимости логарифма среднего значения дисперсии от логарифма величины блоков m разбиения исходной выборки данных. При наличии самоподобия, полученная зависимость должна иметь линейный вид. Таким образом, если аппроксимировать логарифмические зависимости линейной функцией, то с помощью метода наименьших квадратов можно вычислить коэффициенты данного линейного уравнения и коэффициент корреляции данных с линейной зависимостью. Тангенс угла наклона линейной зависимости связан с параметром Харста (Hurst) H.

Важным вопросом является решение задачи выбора параметров информационных процессов, в поведении которых можно определить самоподобие. Анализ данных, описывающих информационное пространство показывает, что можно выделить две взаимосвязанные подсистемы:

- “медленную”, в которой обрабатывается информация, медленно изменяющуюся или почти не изменяющаяся с течением времени (научные, культурные, религиозные, художественные и т.д. знания);
- “быструю”, к которой можно отнести информацию политического, экономического, спортивного и т.д. характера.

Каждая из подсистем содержит определенные наборы кластеров информации, со своими векторами, задающими их положение. Введем понятие директор. *Директором* будем называть условное направление в информационном пространстве, которое определяется взаимным усреднением направлений всех рассматриваемых векторов. Направление директора может быть рассчитано с помощью метода наименьших квадратов для отклонения углов векторов, задающих положение центров информационных кластеров от условного направления, которое принимается за директор. Используя данную методику можно получить “медленный” и “быстрый” директор и при исследовании самоподобия определять зависимость от времени угла между двумя этими директорами. В частности, таким образом можно определить интервал или период самоподобия информационного пространства. Для определения самоподобия процессов, приводящих к реализации интересующего информационного события можно исследовать зависимость от времени углов между каждым из директоров и вектором, описывающим в информационном пространстве данное прогнозируемое событие.

Построение разностных схем вероятностных переходов между состояниями в информационном пространстве

В разрабатываемой нами модели процессов, протекающих в информационном пространстве, величина среднего значения текущего состояния проекций векторов x_i, характеризующих положение центров новостных кластеров в информационном пространстве на ось прогнозируемого события, может случайным образом увеличиваться за счет того, что величина ε больше величины ξ на каждом шаге (или нескольких последовательных шагах), или случайным образом уменьшаться, если величина ε меньше величины ξ. В конечном счете, состояние x_i окажется вблизи *порога прогнозируемого события*, равного величине вектора $\mathbf{X_{bs}}$.

Обозначим все множество состояний, которые описывают возможность появления некоторого события на оси прогнозирования, как X. Состояние, наблюдаемое в момент времени t можно обозначить, как x_i ($x_i \in X$). Интервал времени, за который возможно изменение состояния x_i обозначим как τ_0. В данном случае любое значение текущего времени t=h τ_0, где h – номер шага перехода между состояниями (процесс перехода между состояниями становится квазинепрерывным с бесконечно малым временным интервалом τ_0), h=0,1,2,3, 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 $x_i-\xi$. Величины ε и ξ принадлежат области определения x_i и являются параметрами моделируемых процессов. Кроме того, на $x_i+\varepsilon$, и $x_i-\xi$ необходимо наложить ограничения: $x_i+\varepsilon \le L_1$ (L_1 – верхняя граница множества X) и $x_i-\xi \ge L_2$(L_2 – нижняя граница множества X). В самом простом случае ε и ξ являются для любого шага h постоянными величинами.

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

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

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

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

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

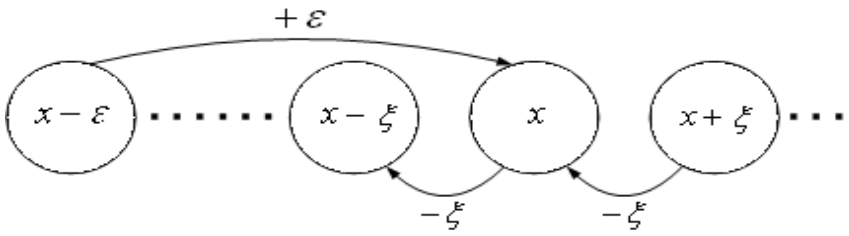

Рис. 1. Схема возможных переходов между состояниями системы (или процесса) на h+1 шаге

Поясним выражение (1) и представленную на рисунке 1 схему. Вероятность перехода в состояние x на шаге h *P(x,h+1)* определяется суммой вероятностей переходов в это состояние из состояний (*x-ε*) - *P(x-ε,h)* и (*x+ξ*) - *P(x+ξ,h)* в которых находилась система на шаге h за вычетом вероятности перехода (*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.

В данном случае мы рассматриваем Марковский непрерывный процесс, в котором система не обладает памятью состояний, однако в реальности в системе, которой является общество, всегда *остается некоторая память* о предыдущем состоянии. Следовательно, предлагаемая модель должна это учитывать. Для этого определим вероятности *P(x-ε,h), P(x+ξ,h)* и *P(x,h*) через состояния на h-1 шаге. Аналогично схеме, представленной на рисунке 1, изобразим схемы соответствующих переходов (см. рис. 2), и учитывая, что ε и ξ являются некоторыми постоянными величинами для любого шага h запишем:

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

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

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

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

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

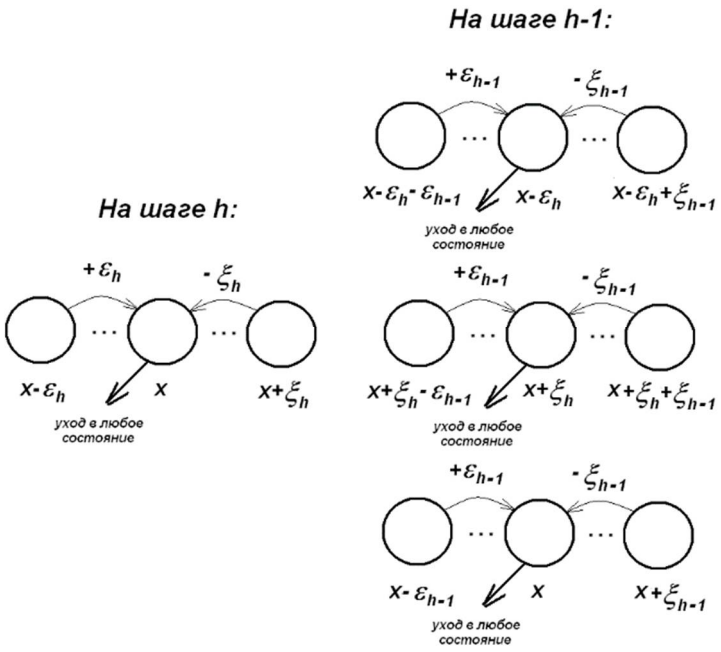

Рис. 2. Схема возможных переходов на h-1 шаге для определения вероятностей P(x-ε,h), P(x+ξ,h) и P(x,h)

Заметим, что в левой части уравнения (5) мы имеем число шагов (h+1), а в правой (h-1). Для того чтобы не проводить разложение правой части уравнения (5) в ряд Тейлора в окрестности числа шагов h (или по времени), а только в окрестности точки x, преобразуем (5) к виду:

$$P(x,h+2)=\{P(x-2\varepsilon,h)+P(x-\varepsilon+\xi,h)-P(x-\varepsilon,h)\}+\{P(x+\xi-\varepsilon,h)+$$
$$+P(x+\xi,h)-P(x+\xi,h)\}-P(x-\varepsilon,h)-P(x+\xi,h-1)+P(x,h). \quad (6)$$

Далее учитывая, что $t=h\cdot\tau_0$, где t – время процесса, h – номер шага, τ_0 – длительность одного шага перейдем от h к t и проведем соответствующие разложения в ряд Тейлора:

$$P(x,h+2)=P(x,t)+2\tau_0\frac{dP(x,t)}{dt}+\frac{(2\tau_0)^2}{2}\frac{d^2P(x,t)}{dt^2}+\cdots$$
$$P(x-2\varepsilon,h)=P(x,t)-2\varepsilon\frac{dP(x,t)}{dx}+\frac{(2\varepsilon)^2}{2}\frac{d^2P(x,t)}{dx^2}+\cdots$$
$$P(x-\varepsilon+\xi,h)=P(x,t)-(\varepsilon-\xi)\frac{dP(x,t)}{dx}+\frac{(\varepsilon-\xi)^2}{2}\frac{d^2P(x,t)}{dx^2}+\cdots$$
$$P(x-\varepsilon,h)=P(x,t)-\varepsilon\frac{dP(x,t)}{dx}+\frac{\varepsilon^2}{2}\frac{d^2P(x,t)}{dx^2}+\cdots$$
$$P(x+\xi-\varepsilon,h)=P(x,t)+(\xi-\varepsilon)\frac{dP(x,t)}{dx}+\frac{(\xi-\varepsilon)^2}{2}\frac{d^2P(x,t)}{dx^2}+\cdots$$
$$P(x+2\xi,h)=P(x,t)+2\xi\frac{dP(x,t)}{dx}+\frac{(2\xi)^2}{2}\frac{d^2P(x,t)}{dx^2}+\cdots$$
$$P(x+\xi,h)=P(x,t)+\xi\frac{dP(x,t)}{dx}+\frac{\xi^2}{2}\frac{d^2P(x,t)}{dx^2}+\cdots$$
$$P(x-\varepsilon,h)=P(x,t)-\varepsilon\frac{dP(x,t)}{dx}+\frac{\varepsilon^2}{2}\frac{d^2P(x,t)}{dx^2}+\cdots$$
$$P(x+\xi,h)=P(x,t)+\xi\frac{dP(x,t)}{dx}+\frac{\xi^2}{2}\frac{d^2P(x,t)}{dx^2}+\cdots$$

Подставив полученные разложения в уравнение (6) находим:

$$2\tau_0\frac{dP(x,t)}{dt}+\frac{(2\tau_0)^2}{2}\frac{d^2P(x,t)}{dt^2}=\{\varepsilon^2+(\varepsilon-\xi)^2+\xi^2\}^2\frac{d^2P(x,t)}{dx^2}+2(\xi-\varepsilon)\frac{dP(x,t)}{dx}.$$

Или в обобщенном виде:

$$\frac{dP(x,t)}{dt}=a\frac{d^2P(x,t)}{dx^2}-b\frac{dP(x,t)}{dx}-c\frac{d^2P(x,t)}{dt^2}, \quad (7)$$

где: $a=\frac{\varepsilon^2-\varepsilon\xi+\xi^2}{\tau_0}$; $b=\frac{\varepsilon-\xi}{\tau_0}$; $c=\tau_0$

Член уравнения вида $\frac{\mathrm{dP(x,t)}}{\mathrm{dx}}$ – описывает упорядоченный переход либо в состояние, когда оно

увеличивается ($\varepsilon > \xi$), либо, когда оно уменьшается ($\varepsilon < \xi$); член уравнения вида $\frac{d^2P(x,t)}{dx^2}$ – описывает случайное изменение состояния (*неопределенность изменения*). Член уравнения вида $\frac{dP(x,t)}{dt}$– можно определить, как скорость общего изменения состояния системы с течением времени; член уравнения вида $\frac{d^2P(x,t)}{dt^2}$ – описывает процесс, при котором состояния сами становятся источниками возникновения других состояний (*самоорганизация* и ускорение как упорядоченных ($\frac{dP(x,t)}{dx}$) и случайных ($\frac{d^2P(x,t)}{dx^2}$) переходов).

С точки зрения области применимости модели в уравнениях (7) и (8) необходимо учесть ограничение, накладываемое на коэффициент $a=(\varepsilon^2-\varepsilon\xi+\xi^2)/\tau_0$ перед второй производной по x, которая учитывает возможность случайного изменения состояния. Должно выполняться условие $(\varepsilon^2-\varepsilon\xi+\xi^2)\geq(l-x_0)^2$, смысл которого заключается в том, что переход из начального состояния x_0 через порог достижения события не может произойти быстрее, чем за время одного шага τ_0. Если $(\varepsilon^2-\varepsilon\xi+\xi^2)<(l-x_0)^2$, то система переходит через порог достижения события за один шаг.

Формулировка и решение краевой задачи при прогнозировании новостных событий в информационном пространстве

Считая функцию $P(x,t)$ непрерывной, можно перейти от вероятности $P(x,t)$ (уравнение (7)) к плотности вероятности $\rho(x,t)=dP(x,t)/dx$ и сформулировать граничную задачу, решение которой и будет описывать процесс перехода между состояниями в информационном пространстве.

$$\frac{d\rho(x,t)}{dt} = a\frac{d^2\rho(x,t)}{dx^2} - b\frac{d\rho(x,t)}{dx} - c\frac{d^2\rho(x,t)}{dt^2}. \quad (8)$$

Первое граничное условие. Первое граничное условие выберем исходя из следующих соображений: состояние $x=0$ определяет полное отсутствие каких бы то ни было, протекающих в информационном пространстве, процессов, с соответствующими им измеряемыми параметрами. Сама вероятность обнаружить такое состояние может быть отлична от 0 (хотя и должна быть близка к нулю), однако плотность вероятности, определяющую поток в состоянии $x=0$, необходимо положить равной *0* (состояния системы не могут выходить в область отрицательных значений (реализуется условие отражения)), т.е.:

$$\boldsymbol{\rho(x,t)_{x=0} = 0} \ . \quad (a)$$

Второе граничное условие. Рассмотрим состояние информационного пространства с значением величины вектора находящегося вблизи границы возможных значений его состояний обозначим эту предельную величину возможных состояний как L). Сама вероятность обнаружить такое состояние будет отлична от 0. Однако плотность вероятности, определяющая поток в состоянии x=L необходимо положить равной *0* (состояния системы не могут выходить в область значений больше, чем максимально возможная величина (реализуется условие отражения от границы)), т.е.:

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

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

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

Так как начальное условие содержит дельта функцию, то решение для $\rho(x,t)$ разбивается на две области при $x>x_0$ и при $x\leq x_0$. Поскольку исходное дифференциальное уравнение содержит вторую производную по времени, то необходимо второе начальное условие. Оно является не столь очевидным, как первое, но в данном случае можно использовать непрерывность функции для любого момента времени. Наличие δ–функции приводит к тому, что решение, оставаясь непрерывным в точке $x=x_0$, испытывает в ней разрыв производной. При решении задачи с помощью методов операционного исчисления возникает необходимость вычисления интеграла $\frac{\partial}{\partial t}\left\{\int_{x_0-0}^{x_0+0} G(x,p)dx\right\}\Big|_{t=0}$, где $G(x,p)$ – изображение $\rho(x,t)$. Поскольку изображение также как и оригинал являются непрерывными, то данный интеграл равен 0, что позволяет не задавать начальное условие дл я $\frac{\partial G(x,t)}{\partial t}\Big|_{t=0}$ в явном виде.

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 $x \geq x_0$

$$\rho_1(x,t) == -\frac{2}{L} e^{-\frac{t}{2\tau_0}} e^{\frac{(x-x_0)(\varepsilon-\xi)}{2(\varepsilon^2-\varepsilon\xi+\xi^2)}} \sum_{n=1}^{\infty} \frac{sin\left(\pi n\frac{x_0}{L}\right) sin\left(\pi n\frac{L-x}{L}\right)}{\cos(\pi n)} \; ch\left(\frac{t}{\tau_0}\sqrt{\frac{\varepsilon\xi}{4(\varepsilon^2-\varepsilon\xi+\xi^2)} - \frac{\pi^2 n^2(\varepsilon^2-\varepsilon\xi+\xi^2)}{L^2}}\right).$$

При $x < x_0$

$$\rho_2(x,t) == -\frac{2}{L} e^{-\frac{t}{2\tau_0}} e^{\frac{(x-x_0)(\varepsilon-\xi)}{2(\varepsilon^2-\varepsilon\xi+\xi^2)}} \sum_{n=1}^{\infty} \frac{sin\left(\pi n\frac{L-x_0}{L}\right) sin\left(\pi n\frac{x}{L}\right)}{\cos(\pi n)} \; ch\left(\frac{t}{\tau_0}\sqrt{\frac{\varepsilon\xi}{4(\varepsilon^2-\varepsilon\xi+\xi^2)} - \frac{\pi^2 n^2(\varepsilon^2-\varepsilon\xi+\xi^2)}{L^2}}\right).$$

Если реализация прогнозируемого события связана с увеличением величины исходного состояния системы x_0, то интеграл *P(l,t)*:

$$\boldsymbol{P(l,t)} = \int_0^{x_0} \boldsymbol{\rho_2(x,t)dx} + \int_{x_0}^{l} \boldsymbol{\rho_1(x,t)dx}. \quad (9)$$

будет задавать вероятность того, что состояние системы к моменту времени *t* находится на отрезке от *0* до *l* (*l*=$\mathbf{X_{bs}}$), т.е. *порог события l* не будет достигнут.

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

$$Q(l,t) = 1 - P(l,t). \quad (10)$$

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→∞ выполняется условие $Q(l,t)$→1 ($P(l,t)$→0).

Если реализация прогнозируемого события связана с уменьшением величины исходного состояния системы x_0 то интеграл *P(L,t)*:

$$P(l,t) = \int_l^{x_0} \rho_2(x,t)dx + \int_{x_0}^{\infty} \rho_1(x,t)dx \quad (11)$$

будет задавать вероятность того, что состояние системы к моменту времени *t* находится на числовой прямой от *l*<x_0<∞, т.е. *порог события l* не будет достигнут. Вероятность $Q_i(t)$ того, что *порог события l* окажется к моменту времени *t* достигнутым также определяется по формуле (10).

Согласно нашей модели, стохастическая динамика описывается изменением состояния за счет параметров ε и ξ, величины которых определяются множеством различных факторов.

Анализ решения краевой задачи при прогнозировании редких событий в информационном пространстве. Реализация памяти и самоорганизация

Проанализируем полученные результаты. Для моделирования процесса будем считать, что начальное (в момент начала наблюдения) значение величины вектора состояния системы (информационного пространства) равно x_0 (x_0=0,05 –условно принятая величина), величину τ_0 примем равной 1 условной единице времени, ε=0,02 и ξ=0,01, *l*=2–условно принятая величина.

Результаты решения уравнения (10) с использованием (9), функций $\rho_1(x,t)$, $\rho_2(x,t)$ и заданным выше набором параметров и различными *порогами событий* (отметим, что в данном случае прогнозируемое событие наблюдается при росте величины вектора состояния системы), выбранных при моделировании, представлены в графическом виде на рисунке 3. Кривая 1 на рисунке 3 построена для порога события равного 0,1; кривая 2 для величины порога события равной 0,2, а кривая 3 для величины порога события равной 0,3.

Ход кривых на рисунке 3, показывает возможность роста вероятности перехода *порога события* практически сразу после начала процесса, что связано с наличием памяти о предыдущих состояниях системы в разработанной нами модели, и возможности описания значительной самоорганизации системы вследствие учета в дифференциальной модели члена уравнения, имеющего вид $\frac{d^2P(x,t)}{dt^2}$. Кривые 1 и 2 на рисунке 3 демонстрируют, что чем ближе значение величины состояния системы x_0 в момент начала наблюдения к *порогу события*, тем быстрее возрастает вероятность перехода (кривая 1 построена для *порога события* равного 0,1, а кривая 2 для 0,2, при одинаковой начальной величине вектора состояния системы 0,05).

Второй особенностью, предлагаемой нами модели, является возможность нескольких скачкообразных изменений вероятности перехода через *порог события*.

Третьей особенностью, предлагаемой нами модели, является наличие волнообразных явлений в поведении величины достижения *порога события*.

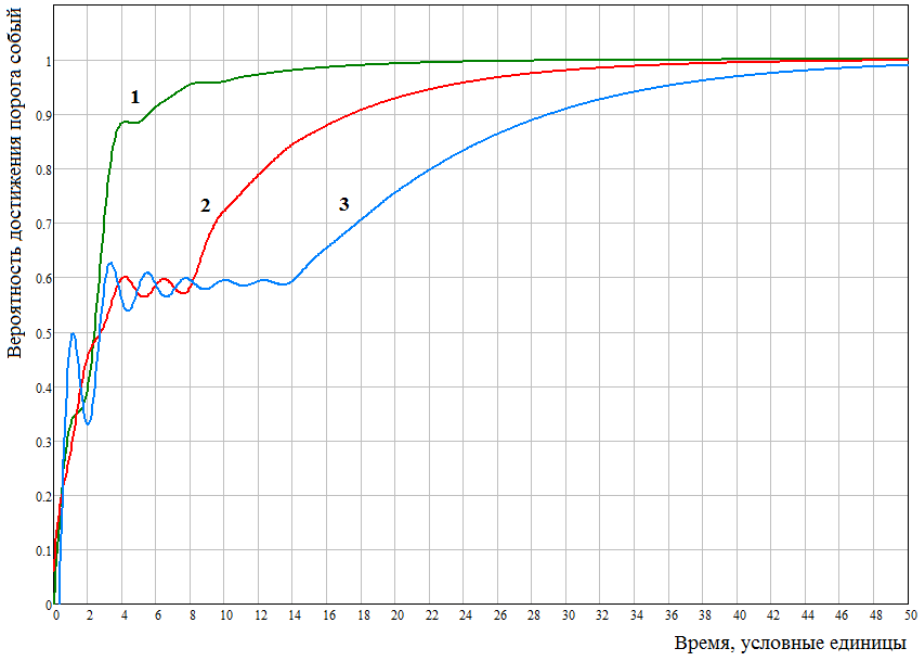

Рис. 3. Графическое представление результатов моделирования преодоления порога событий, при росте величины состояния

Разработанная модель учитывает многие основные свойства редких событий: неопределенность во времени их проявления, стохастичность с неизвестными характеристиками, наличие памяти в системе в которой происходит событие, самоорганизацию информации.

Разработанная нами модель позволяет проанализировать возможность прогнозирования редкого новостного события типа «Чёрный лебедь» во взаимосвязи с различными кластерами в информационном пространстве. Анализ значений времен достижения редких событий может позволить делать их достаточно точный прогноз, при заданном уровне вероятности реализации (например, 0,90 или 0,95).

Необходимо отметить, что каждое редкое новостное событие типа «Чёрный лебедь» имеет плохо структурированных и слабо выраженных предшественников, появление которых не было обнаружено, но они являются значимыми индикаторами, на основе которых в ретроспективе, как правило, дается рационалистическое объяснение произошедшему явлению. Исследование самоподобия информационных процессов может позволить определить период или интервал самоподобия, что является важным с прогностической точки зрения.

Методика экспериментальной проверки предлагаемых моделей. Ресурсоемкость исследований и алгоритм реализации

Экспериментальная проверка разработанных нами моделей может быть проведена на большом массиве текстовых документов с использованием следующего алгоритма:

a) Проводим представление текстовых документов в информационном пространстве с помощью методов математической лингвистики к векторной форме, и осуществляем в какой-то момент времени t=0 кластеризацию по различным смысловым группам, используя алгоритм k-средних, Scatter-Gather, BIRCH или алгоритмы иерархической кластеризации. Выделяем быстро и медленно изменяющиеся подсистемы данных в информационном пространстве, и определяем для них направления директоров;
b) Задаем вектор события в информационном пространстве посредством его текстового описания. Величина вектора прогнозируемого события задает на оси этого события порог его достижения (***I***). Определяем углы между директорами и вектором прогнозируемого события;
c) Проводим отображение векторов, определяющих положение центров новостных кластеров на направление оси, задаваемой вектором прогнозируемого события момент времени t=0 и находим их среднее значение x_0;
d) Спустя некоторый интервал времени измерения (назовем его τ_0) определяем новые вектора, задающие положение центров новостных кластеров и их отображения на ось прогнозируемого события. Разделяем отображения на две группы. В первой будут все вектора, для которых значения отображений увеличились, во второй для которых уменьшились, по сравнению с предыдущими значениями. Находим средние значения по группам и определяем значения

величин трендов увеличения (ε) и уменьшения (ξ) по отношению к начальному состоянию x_0. Величины ε и ξ определяются разностью текущего среднего значения по группе и предыдущего общего состояния x_0. Выделяем быстро и медленно изменяющиеся подсистемы данных в информационном пространстве, определяем для них новые направления директоров и углы между директорами и вектором прогнозируемого события;

e) Используя величины x_0, ε, ξ, τ_0, l и выражения 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 моделируем достижение порога события, и определяем время его достижения (при заданной вероятности реализации, например 0,9 или 0,8);

f) Через новый интервал времени измерения τ_0 повторяем процедуры, описанные в пунктах b) – e). По полученной зависимости от времени углов между директорами и вектором прогнозируемых событий определяем наличие или отсутствие самоподобия в процессах, протекающих в информационном пространстве, и при его наличии определяем параметры самоподобия (период или интервал, что важно для прогнозирования интересующего события).

Представление текстовых данных в векторной форме с помощью методов математической лингвистики создает векторное пространство размерности $\mathbf{R^m}$, где m - может достигать величины нескольких десятков миллионов. Использование для проведения исследований больших коллекций текстовых документов является в техническом плане чрезвычайно ресурсоемкой задачей, решение которой связано с обработкой Экзобайтных объемов текстовой информации, и требует привлечения вычислительных кластеров с Петафлопной производительностью. Поэтому для предварительной проверки модели можно ограничиться меньшей выборкой данных. Например, использовать существующие общедоступные базы текстовой научной информации, и как один из возможных вариантов взять информацию из Wikipedia, научных и новостных порталов, публикующих новости науки и техники. В качестве редкого события можно описать, например, информационную карту любого из футуристических научных прогнозов, сделанных техническим директором компании Google Рэем Курцвейлом (например, персональные компьютеры достигнут вычислительной мощности мозга человека, компьютер сможет пройти тест Тьюринга, 3D – принтеры начнут печатать человеческие органы, появится небиологический интеллект и т.д.) и оценить время его реализации при разной вероятности.

Выводы

1. Предлагаемая модель учитывает неопределенности в процессе возникновения событий в информационном пространстве и не основывается на статистических характеристиках с заранее предполагаемым законом распределения.
2. Предлагаемая модель показывает возможность роста вероятности перехода порога достижения новостного события в информационном пространстве практически сразу после начала процесса его развития, что связано с учетом памяти о предыдущих состояниях системы и возможности описания значительной самоорганизации вследствие учета в дифференциальной модели второй производной по времени.
3. Предлагаемая модель показывает возможность скачкообразных изменений для вероятности перехода через порог событий и учитывает наличие в её поведении волнообразных явлений.
4. Разработанная модель позволяет создать алгоритм анализа взаимосвязи кластеров информации в информационном пространстве с возможностью реализации любого события, а также определить прогнозируемое время его возникновения.

Литература

1. Nassim N.T. The Black Swan: The Impact of the Highly Improbable. Random House.– 2007. – P. 401.
2. Джон Дж. Мерфи Технический анализ фьючерсных рынков: теория и практика. – М.: Сокол, 1996. – 592 с.
3. Кендал М. Дж., Стьюарт А. Многомерный статистический анализ и временные ряды. – М.: Наука, 1976. – 736 с.
4. Автухович Э.В. Бурова Н.К., Дорин Б.Л., Панов С.С., Петров А.А., Поспелов И.Г., Поспелова И.И., Ташлицкая Я.М., Чуканов С.В., Шананин А.А., Шапошник Д.В. Оценка потенциала роста экономики России с помощью математической модели. – М.: ВЦ РАН, 2000. – 154 с.
5. Петров А.А., Поспелов И.Г., Поспелова И.И., Хохлов М.А., Шипулина Г.Е. Новые принципы и методы разработки макромоделей экономики и модель современной экономики России. – М.: ВЦ РАН, 2006. – 242 с.
6. Петров А.А., Поспелов И.Г. Математические модели экономики России // Вестник РАН, Т.79, № 6, 2009. – С. 492-506.
7. Алескеров Ф. Т. «Простые и простейшие полупорядки». // Доклады РАН, т.387, № 2, 2002 – С. 175—177.
8. Алескеров Ф. Т. Индексы влияния, учитывающие предпочтения участников по созданию коалиций, Доклады РАН. – 2007. – т. 414, № 5. – P. 594—597.
9. Yingyuan Xiao, Pengqiang Ai, Ching-hsien Hsu, Hongya Wang, Xu Jiao. Time-ordered Collaborative Filtering for News Recommendation. // China Communication. – Vol. 12, № 12. – P. 53-62.

10. Daqiang Zhang, Ching-Hsien Hsu, Min Chen, Quan Chen, Naixue Xiong, Jaime Lloret. Cold-Start Recommendation Using Bi-Clustering and Fusion for Large-Scale Social Recommender Systems. // IEEE Transactions on Emerging Topics in Computing. – 2014. – Vol. 2, № 2. – P.239-250.
11. Asela Gunawardana, Christopher Meek, Puyang Xu. A Model for Temporal Dependencies in Event Streams. / Microsoft Research.– 2011. – P. 1-8.
12. Zhukov D.O., Lesko S.A. Trends, self-similarity and forecasting of news events in the information domain, its structure and director. // IEEE International Conference on Social Communication. - 2015, - P. 870-873.
13. Feldman R., Sanger J. The Text Mining Handbok. Cambridge: Cambridge University Press. – 2007.
14. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. Introduction to Information Retrieval. Cambridge University Press, 2008. – P. 544.
15. Pantel, P., Lin, D. Discovering word senses from text. In Proceedings of the Eighth ACM SIGKDD. // International Conference on Knowledge Discovery and Data Mining, 2002. – P. 613–619, Edmonton, Canada.
16. Turney, P. D. The latent relation mapping engine: Algorithm and experiments. // Journal of Artificial Intelligence Research, 33, 2008. – P. 615–655.
17. Clegg R.G. A practical guide to measuring the Hurst parameter. // Computing science technical report. – 2005. – № CS–TR–916. – P. 125-138.

References

1. Nassim N.T. The Black Swan: The Impact of the Highly Improbable. Random House.– 2007. – P. 401.
2. Dzhon Dzh. Merfi Tekhnicheskiy analiz fyuchersnykh rynkov: teoriya i praktika. – M.: Sokol, 1996. –592 s.
3. Kendal M. Dzh., Styuart A. Mnogomernyy statisticheskiy analiz i vremyannye ryady. – M.: Nauka, 1976. –736 s.
4. Avtukhovich E.V. Burova N.K., Dorin B.L., Panov S.S., Petrov A.A., Pospelov I.G., Pospelova I.I., Tashliczkaya Ya.M., ChukanovS.V., Shananin A.A., Shaposhnik D.V. Ocenka potenciala rosta ekonomiki Rossii s pomoshchyu matematicheskoy modeli. – M.: VC RAN, 2000. –154 s.
5. Petrov A.A., Pospelov I.G., Pospelova I.I., Khokhlov M.A., Shipulina G.E. Novye principy i metody razrabotki makromodeley ekonomiki i model sovremennoy ekonomiki Rossii. – M.: VC RAN, 2006. – 242 s.
6. Petrov A.A., Pospelov I.G. Matematicheskie modeli ekonomiki Rossii// Vestnik RAN, T.79, № 6, 2009. – S. 492-506.
7. Aleskerov F. T. «Prostye i prosteyshie poluporyadki»// Doklady RAN, т.387, № 2, 2002 – S. 175—177.
8. Aleskerov F. T. Indeksy vliyaniya, uchityvayushchie predpochteniya uchastnikov po sozdaniyu koaliciy. –Doklady RAN. – 2007. – т. 414, № 5. –S. 594—597.
9. Yingyuan Xiao, Pengqiang Ai, Ching-hsien Hsu, Hongya Wang, Xu Jiao. Time-ordered Collaborative Filtering for News Recommendation. // China Communication. – Vol. 12, № 12. – P. 53-62.
10. Daqiang Zhang, Ching-Hsien Hsu, Min Chen, Quan Chen, Naixue Xiong, Jaime Lloret. Cold-Start Recommendation Using Bi-Clustering and Fusion for Large-Scale Social Recommender Systems. // IEEE Transactions on Emerging Topics in Computing. – 2014. – Vol. 2, № 2. – P.239-250.
11. Asela Gunawardana, Christopher Meek, Puyang Xu. A Model for Temporal Dependencies in Event Streams. / Microsoft Research.– 2011. – P. 1-8.
12. Zhukov D.O., Lesko S.A. Trends, self-similarity and forecasting of news events in the information domain, its structure and director. // IEEE International Conference on Social Communication. - 2015, P. 870-873.
13. Feldman R., Sanger J. The Text Mining Handbok. Cambridge: Cambridge University Press. – 2007.
14. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. Introduction to Information Retrieval. Cambridge University Press, 2008. – P. 544.
15. Pantel, P., Lin, D. Discovering word senses from text. In Proceedings of the Eighth ACM SIGKDD. // International Conference on Knowledge Discovery and Data Mining, 2002. – P. 613–619, Edmonton, Canada.
16. Turney, P. D. The latent relation mapping engine: Algorithm and experiments. // Journal of Artificial Intelligence Research, 33, 2008. – P. 615–655.
17. Clegg R.G. A practical guide to measuring the Hurst parameter. // Computing science technical report. – 2005. – № CS–TR–916. – P. 125-138.

Поступила: 20.10.2016

Об авторах:

Сигов Александр Сергеевич, президент Московского технологического университета (МИРЭА), академик РАН, заведующий кафедрой наноэлектроники физико-технологического института Московского технологического университета, доктор технических наук;

Жуков Дмитрий Олегович, профессор кафедры информационного противоборства института комплексной безопасности и специального приборостроения Московского технологического университета, доктор технических наук, zhukovdm@yandex.ru;

Новикова Ольга Александровна, заместитель заведующего аспирантурой Московского технологического университета, ol-novikova@bk.ru.