Информационные процессы, Том 18, № 2, 2018, стр. 106–117

МАТЕМАТИЧЕСКИЕ МОДЕЛИ, ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ

Модель прогнозирования информационных событий на основе решения краевой задачи для систем с реализацией памяти и самоорганизацией[1]

А.С. Сигов*, Д.О. Жуков*, Т.Ю. Хватова, Е.Г. Андрианова***

**Московский технологический университет (МИРЭА), Москва, Россия*
***Санкт-Петербургский политехнический университет Петра Великого (СПбПУ), Санкт-Петербург, Россия*

Поступила в редколлегию 21.05.2018

Аннотация—Одной из проблем прогнозирования новостных событий является разработка моделей, позволяющих работать со слабоструктурированным информационным пространством текстовых документов. Отличительной особенностью такого новостного пространства является стохастический характер протекающих в нем процессов, наличие памяти и возможность самоорганизации информации. Представляется интересным создание модели прогнозирования событий на основе стохастической динамики изменения образов (или состояния информационного пространства) новостных кластеров, учитывающей память и самоорганизацию. В статье рассматриваются схемы вероятностей переходов между состояниями в информационном пространстве, на основании чего выводится нелинейное дифференциальное уравнение второго порядка, формулируется и решается краевая задача для прогнозирования новостных событий. Анализ описанной в статье модели показывает возможность роста вероятности достижения прогнозируемого события, практически сразу после начала процесса изменения структуры новостных кластеров, наличие в вероятности достижения события резких скачков и осцилляций.

КЛЮЧЕВЫЕ СЛОВА: стохастическая динамика изменения состояний информационной системы, самоорганизация, процессы с учетом памяти, порог новостного события, информационное пространство, новостной кластер, кластеризация новостей

1. ВВЕДЕНИЕ

Возьмём коллекцию из N текстовых документов, описывающих новостные события за некоторый период времени. Используя методы математической лингвистики [1–4], создадим с помощью словаря терминов размера – M, векторное представление текстов в информационном пространстве (размерность которого будет R^M). Каждому документу коллекции можно поставить в соответствие вектор $X_i = \{x_{1,i}, x_{2,i}, \ldots, x_{k,i}, \ldots, x_{M,i}\}$, где i – принимает значения от 1 до N, а каждый элемент вектора $x_{k,i}$ характеризует нормированную частоту вхождений терминов словаря в документ. Вектора X_i образуют матрицу размера N на M: термин –

[1] Работа выполнена за счет финансирования Министерством образования и науки РФ конкурсной части государственных заданий высшим учебным заведениям и научным организациям по выполнению инициативных научных проектов, № 28.2635.2017/ПЧ.

документ:

$$\begin{vmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,i} & \dots & x_{1,N} \\ x_{2,1} & x_{2,2} & \dots & x_{2,i} & \dots & x_{2,N} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{j,1} & x_{i,2} & \dots & x_{k,i} & \dots & x_{i,N} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{M,1} & x_{M,2} & \dots & x_{M,i} & \dots & x_{M,N} \end{vmatrix}. \tag{1}$$

Далее с использованием стандартных методов [1–4] проведем кластеризацию текстовых документов (разделение по смысловым группам), используя их векторное представление. 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.

Создадим текстовое описание образа новостного события, для которого необходимо определить вероятность его реализации с течением времени (прогноз). Далее векторизуем текстовое описание прогнозируемого события.

Определим, для какого либо момента времени, проекции векторов центроидов на направление вектора прогнозируемого события (ось прогнозируемого события) и их среднее значение. Величина среднего значения проекций центроидов в данный момент времени будет являться точкой на числовой оси прогнозируемого события, и вследствие изменения структуры кластеров с течением времени, она будет совершать на ней случайные перемещения. Текущую величину среднего значения проекций назовем состоянием информационной системы в данный момент времени. Достижение величины среднего значения проекций центроидов, точки на оси прогнозируемого события, отвечающей величине вектора прогнозируемого события, будем рассматривать как его реализацию [5, 6].

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

2. ПОСТРОЕНИЕ РАЗНОСТНЫХ СХЕМ ВЕРОЯТНОСТЕЙ ПЕРЕХОДОВ МЕЖДУ СОСТОЯНИЯМИ В ИНФОРМАЦИОННОМ ПРОСТРАНСТВЕ. ВЫВОД ОСНОВНОГО УРАВНЕНИЯ МОДЕЛИ

Обозначим величину среднего значения текущего состояния проекций векторов центроидов на ось прогнозируемого события, как x_i (назовем это состоянием 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 окажется вблизи порога прогнозируемого события – l, равного величине вектора X_{bs}.

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), $h = 0, 1, 2, 3, N$. 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 $(x_i + \xi)$.

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

- $P(x-\varepsilon, h)$ – вероятность того, что она находится в состоянии $(x-\varepsilon)$;
- $P(x, h)$ – вероятность того, что она находится в состоянии x;
- $P(x+\xi, h)$ – 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 $P(x, h+1)$ – того, что на следующем $(h+1)$ шаге система окажется в состоянии x, будет определяться несколькими переходами (см. рис. 1):

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

Поясним выражение (2) 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 $(x-\varepsilon)$: $P(x-\varepsilon, h)$ и $P(x+\xi)$: $P(x+\xi, h)$ 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.

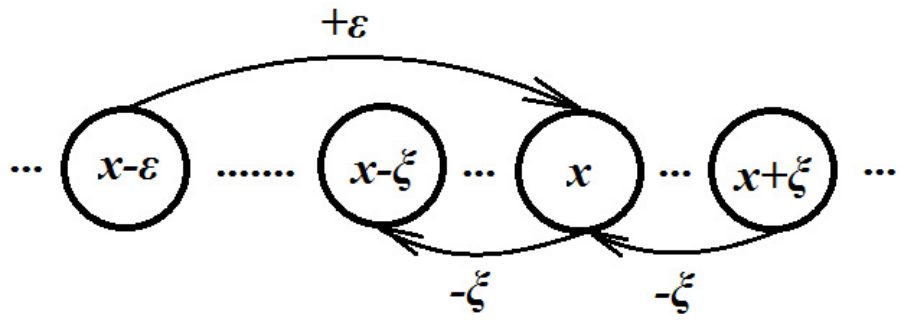

Рис. 1. Схема возможных переходов между состояниями системы на $h+1$ шаге

В данном случае мы рассматриваем Марковский непрерывный процесс, в котором система не обладает памятью состояний, однако в реальности в информационном пространстве может сохраняться память о предыдущем состоянии. Для учета памяти определим вероятности $P(x-\varepsilon, h)$, $P(x+\xi, h)$ 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:

$$P(x-\varepsilon, h) = P(x-2\varepsilon, h-1) + P(x-\varepsilon+\xi, h-1) - P(x-\varepsilon, h-1), \tag{3}$$

$$P(x+\xi, h) = P(x+\xi-\varepsilon, h-1) + P(x+2\xi, h-1) - P(x+\xi, h-1), \tag{4}$$

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

Подставив (3), (4) и (5) в уравнение (2) получим:

$$\begin{aligned} P(x, h+1) = {} & \{P(x-2\varepsilon, h-1) + P(x-\varepsilon+\xi, h-1) - P(x-\varepsilon, h-1)\} + \\ & + \{P(x+\xi-\varepsilon, h-1) + P(x+2\xi, h-1) - P(x+\xi, h-1)\} - \\ & - P(x-\varepsilon, h-1) - P(x+\xi, h-1) + P(x, h-1) \end{aligned} \tag{6}$$

Заметим, что в левой части уравнения (6) мы имеем число шагов $(h+1)$, а в правой $(h-1)$. Для того чтобы не проводить разложение правой части уравнения (6) в ряд Тейлора в

окрестности числа шагов h (или по времени), а только в окрестности точки x, преобразуем (6) к виду:

$$\begin{aligned} P(x,h+2) = {} & \{P(x-2\varepsilon,h) + P(x-\varepsilon+\xi,h) - P(x-\varepsilon,h)\}+ \\ & + \{P(x+\xi-\varepsilon,h) + P(x+\xi,h) - P(x+\xi,h)\}- \\ & - P(x-\varepsilon,h) - P(x+\xi,h-1) + P(x,h). \end{aligned} \tag{7}$$

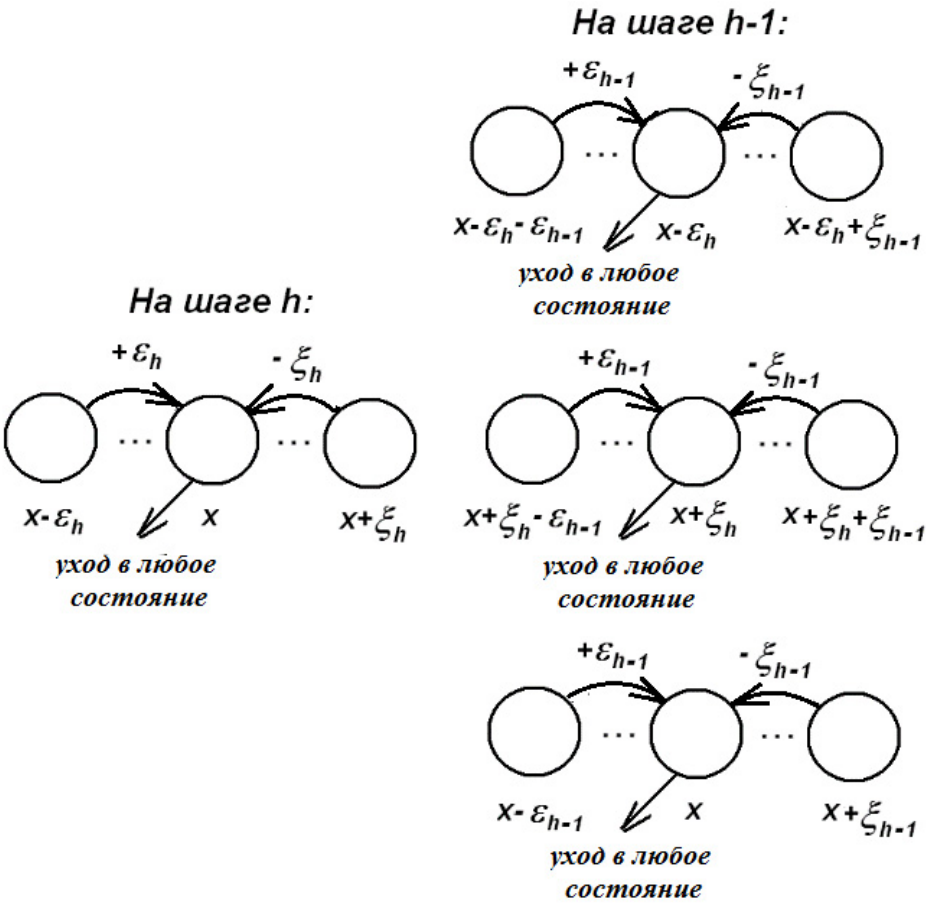

Рис. 2. Схема возможных переходов на $h-1$ шаге, для определения вероятностей $P(x-\varepsilon,h)$, $P(x+\xi,h)$ и $P(x,h)$

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

$$P(x,h+2) = P(x,t) + 2\tau\frac{dP(x,t)}{dt} + \frac{(2\tau)^2}{2}\frac{d^2P(x,t)}{dt^2} + \cdots,$$

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

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

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

Подставив полученные разложения в уравнение (7) находим:

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

Или в обобщенном виде:

$$\frac{dP(x,t)}{dt}=a\frac{d^2P(x,t)}{dx^2}-b\frac{dP(x,t)}{dx}-c\frac{d^2P(x,t)}{dt^2} \tag{8}$$

где:

$$a=\frac{\varepsilon^2-\varepsilon\xi+\xi^2}{\tau};\quad b=\frac{\varepsilon-\xi}{\tau};\quad c=\tau.$$

Член уравнения вида $\frac{dP(x,t)}{dx}$ – 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 $\left(\frac{dP(x,t)}{dx}\right)$ и случайных $\left(\frac{d^2P(x,t)}{dx^2}\right)$ переходов).

С точки зрения области применимости модели, в уравнении (8) необходимо учесть ограничение, накладываемое на коэффициент $a=(\varepsilon^2-\varepsilon\xi+\xi^2)/\tau$ 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 ($\boldsymbol{l}$) не может произойти быстрее, чем за время одного шага τ. Если $(\varepsilon^2-\varepsilon\xi+xi^2)<(l-x_0)^2$, то система переходит через порог достижения события за один шаг.

3. ПОСТРОЕНИЕ РАЗНОСТНЫХ СХЕМ ВЕРОЯТНОСТЕЙ ПЕРЕХОДОВ МЕЖДУ СОСТОЯНИЯМИ В ИНФОРМАЦИОННОМ ПРОСТРАНСТВЕ. ВЫВОД ОСНОВНОГО УРАВНЕНИЯ МОДЕЛИ

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 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, определяющую поток в состоянии $x = 0$, необходимо положить равной 0 (состояния системы не могут выходить в область отрицательных значений (реализуется условие отражения)), т.е.:

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

Второе краевое условие. Ограничим область возможных состояний информационной системы некоторой величиной L и выберем второе краевое условие для состояния $x = L$. Вероятность обнаружить такое состояние с течением времени 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.:

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

В данном случае можно рассмотреть и условие обращения потока в 0 на бесконечности:

$$\rho(x,t)_{x=\infty} = 0.$$

Уравнение (8) содержит вторую производную по времени и для формулировки краевой задачи необходимо задать два начальных условия. Поскольку в момент времени $t = 0$ состояние системы уже может быть равно некоторому значению x_0, то первое 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 $x \leq x_0$. Второе начальное условие 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 от изображения

$$\frac{\partial}{\partial t}\left\{\int_{x_0-0}^{x_0+0} \overline{G}(x,p)dx\right\}\Bigg|_t = 0,$$

где $\overline{G}(x,p)$ – изображение $\rho(x,t)$. Поскольку изображение также как и оригинал являются непрерывными, то данный интеграл равен 0, что позволяет не задавать начальное условие для $\frac{\partial \overline{G}(x,t)}{\partial t}\Big|_t = 0$ в явном виде.

Используя методы операционного исчисления, а также начальные и краевые условия (a) и (b) 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 $x \geq 0$

$$\begin{aligned} p_1(x,t) = -\frac{2}{L} e^{-\frac{t}{2\tau}} e^{\frac{(x-x_0)(\varepsilon-\xi)}{2(\varepsilon^2-\varepsilon\xi+\xi^2)}} \sum_{n=1}^{\infty} \frac{\sin\left(\pi n \frac{x_0}{L}\right)\sin\left(\pi n \frac{L-x}{L}\right)}{\cos(\pi n)} \\ \times \operatorname{ch}\left(\frac{t}{\tau}\sqrt{\frac{\varepsilon\xi}{4(\varepsilon^2-\varepsilon\xi+\xi^2)} - \frac{\pi^2 n^2(\varepsilon^2-\varepsilon\xi+\xi^2)}{L^2}}\right). \end{aligned} \tag{9a}$$

При $x < 0$

$$p_2(x,t) = -\frac{2}{L} e^{-\frac{t}{2\tau}} e^{\frac{(x-x_0)(\varepsilon-\xi)}{2(\varepsilon^2-\varepsilon\xi+\xi^2)}} \sum_{n=1}^{\infty} \frac{\sin\left(\pi n \frac{L-x_0}{L}\right) \sin\left(\pi n \frac{x}{L}\right)}{\cos(\pi n)} \times \operatorname{ch}\left(\frac{t}{\tau}\sqrt{\frac{\varepsilon\xi}{4(\varepsilon^2-\varepsilon\xi+\xi^2)} - \frac{\pi^2 n^2(\varepsilon^2-\varepsilon\xi+\xi^2)}{L^2}}\right). \tag{9b}$$

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

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

будет задавать вероятность того, что состояние системы к моменту времени t находится на отрезке от 0 до l ($l = X_{bs}$), 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:

$$Q(l,t) = 1 - P(l,t). \tag{10}$$

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 $t \to \infty$ выполняется условие $Q(l,t) \to 1$ $(P(l,t) \to 0)$.

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

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

будет задавать вероятность того, что состояние системы к моменту времени t находится на числовой прямой от $l < x_0 < \infty$, 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).

4. ОПРЕДЕЛЕНИЕ ПАРАМЕТРОВ МОДЕЛИ ПРОГНОЗИРОВАНИЯ СОБЫТИЙ НА ОСНОВЕ ИЗМЕНЕНИЯ СТРУКТУРЫ КЛАСТЕРОВ В ИНФОРМАЦИОННОМ ПРОСТРАНСТВЕ

Модель прогнозирования информационных событий 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 её основе.

В результате кластеризации N текстовых документов по смысловым группам, из матрицы (1) можно выбрать подмножества векторов, каждое из которых образует свой кластер (обозначим общее число кластеров – W). В свою очередь, внутри каждого кластера W, разделяем вектора на привязанные к определенной дате времени новые подмножества.

Для каждой даты времени, внутри каждого кластера, для выбранного подмножества векторов определяем координаты центроида:

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

где $c(t)_{k,j}$ – среднее арифметическое координат векторов входящих в данный кластер для момента времени t, а j принимает значения от 1 до W. Затем даем текстовое описание прогнозируемого события и создаем его вектор. Находим для данного момента времени t величины проекций векторов $C_j(t)$ на направление оси, заданной вектором текстового описания прогнозируемого события (обозначим проекции, как $S_j(t)$). Спустя интервал времени τ выбираем в каждом кластере подмножество векторов, соответствующее дате времени $(t+\tau)$ и определяем новые координаты центроидов:

$$C_j(t+\tau) = \{c(t+\tau)_{1,j}, c(t+\tau)_{2,j}, \dots c(t+\tau)_{k,j}, \dots c(t+\tau)_{M,j}\}$$

и новые значения проекций $S_j(t+\tau)$. Находим для проекций отклонения (обозначим их, как $\Delta S_j(\tau)$) за интервал времени τ: $\Delta S_j(t+\tau) - S_j(t)$ и сортируем их на две группы: $\Delta S_j(\tau) < 0$ и $\Delta S_j(\tau) > 0$. Среднее значение для проекций отклонения по группе $\Delta S_j(\tau) < 0$ примем за величину тренда уменьшения ξ, а по группе $\Delta S_j(\tau) > 0$ за величину тренда увеличения ε. Величины трендов уменьшения ξ и увеличения ε, можно определять не только за один интервал времени τ, но и за несколько.

5. АНАЛИЗ МОДЕЛИ ПРОГНОЗИРОВАНИЯ СОБЫТИЙ НА ОСНОВЕ РЕШЕНИЯ КРАЕВОЙ ЗАДАЧИ ДЛЯ СИСТЕМ С РЕАЛИЗАЦИЕЙ ПАМЯТИ И САМООРГАНИЗАЦИИ

Проанализируем полученную модель. 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, $\varepsilon = 0,02$ и $\xi = 0,01$, $l = 2$ – условно принятая величина.

Для примера анализа модели примем среднее значение величин проекций векторов центроидов кластеров на ось прогнозируемого события равным $0,05$ условных единиц ($x_0 = 0,05$), величину интервала времени τ примем равной 1 условной единице, величины $\varepsilon = 0,02$ и $\xi = 0,01$, $l = 2$.

Результаты решения уравнения (10) с использованием (9), функций $\rho_1(x,t)$, $\rho_2(x,t)$ и заданным выше набором параметров и различными значениями величин порогов событий (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 $0,15$, кривая 3 для величины порога события равной $0,20$, а кривая 4 – $0,25$.

Кривые 1 и 2 на рисунке 3 показывают, что чем ближе значение величины начального состояния информационной системы x_0 в момент времени $t = 0$ начала наблюдения к порогу события, тем быстрее возрастает вероятность перехода. Кривая 1 построена для порога события равного $0,1$, а кривая 2 для $0,15$, при одинаковой начальной величине вектора состояния системы $0,05$. Полученные при моделировании результаты показывают, что чем ближе величина начального состояния системы x_0 к величине порога реализации события, тем быстрее вероятность его достижения приближается к единице.

Кривая 4 на рисунке 3 показывает, что при большой разности между величиной порога события и начального состояния системы x_0 вероятность его достижения имеет осциллирующий характер, при этом она сначала снижается с течением времени, а затем показывает осциллирующий рост. Причем чем дальше значение величины начального состояния системы x_0 от порога события, тем сильнее проявляются осцилляции. Кроме того, существует и отличная от нуля вероятность реализации прогнозируемого события в начальный момент времени

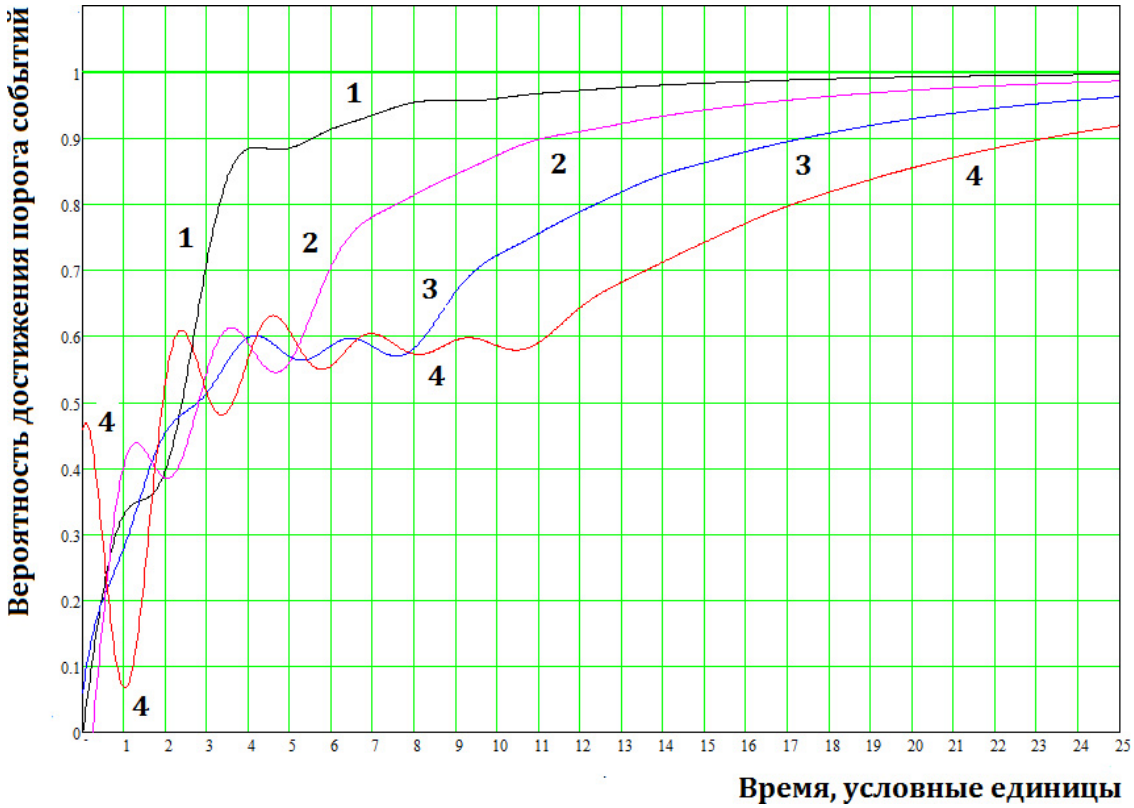

Рис. 3. Результаты моделирования преодоления порога событий, при росте величины состояния (при величине $x_0 = 0,05$; уменьшения за единицу времени τ на величину $\xi = 0,01$ и увеличения на $\varepsilon = 0,02$)

(мгновенная реализация), что косвенно согласуется с тем, что в реализации событий имеется неопределенность и нечеткость.

Рост вероятности перехода через порог события имеет ступенчатый характер, а протяженность ступени во времени зависит от того насколько начальная величина состояния системы x_0 близка к порогу события.

Ход кривых на рисунке 3, показывает возможность роста вероятности достижения порога события, практически сразу после начала процесса. Вероятность перехода через порог события отлична от нуля уже после первого шага, и нелинейно возрастает с течением времени. Это является следствием того, что не только величины ε и ξ определяют изменение состояния x, но и сами состояния x являются источником изменения, вследствие наличия памяти о предыдущих состояниях и самоорганизации, за которую отвечает в дифференциальном уравнении (8) член $\frac{d^2P(x,t)}{dt^2}$.

Результаты анализа модели, представленные на рисунке 3, показывают возможность самоорганизации системы, заключающейся в следующем. Величины изменения состояния x на одном шаге ε (увеличение x) и ξ (уменьшение x) являются сами по себе случайными. Простой арифметический подход говорит о том, что число шагов (обозначим его как q) за которое мог бы быть достигнут порог реализации прогнозируемого события l, не может быть меньше чем $q = (l - x_0)/(\varepsilon - \xi)$. Например, для порогов $l = 0,1;\ 0,2$; и начального состояния $x_0 = 0,05$, при $\varepsilon = 0,02$ и $\xi = 0,01$ для q получим соответственно 5 и 15.

Процесс достижения порога события имеет протяженное во времени плато, величина которого (в единицах вероятности) зависит от начальной величины состояния системы.

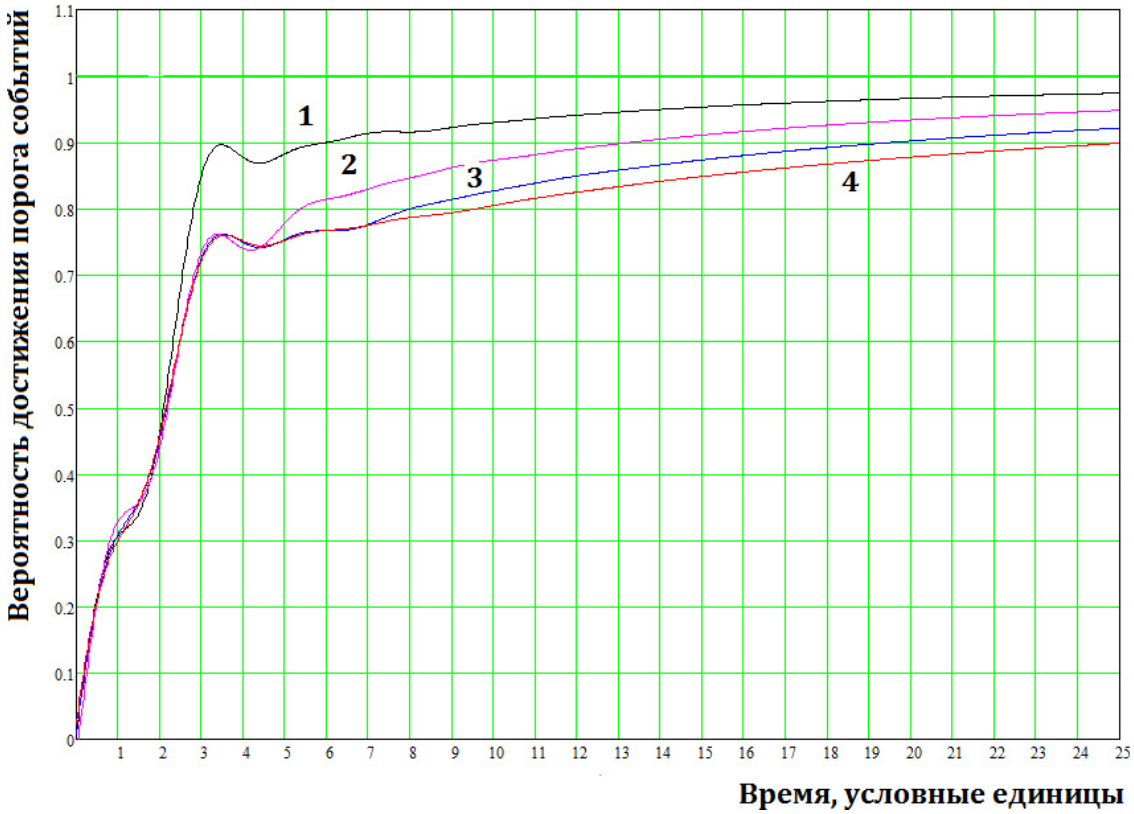

Рис. 4. Результаты моделирования достижения порога прогнозируемого события (при величине $x_0 = 0,05$; уменьшения за единицу времени τ на величину $\xi = 0,02$ и увеличения на $\varepsilon = 0,02$)

При равенстве величин ε и ξ (например ($\varepsilon = \xi = 0,02$) характер хода кривых, описывающих вероятность достижения порога перколяции изменяется (см. рис. 4). В частности, не наблюдается протяженное во времени плато с последующим плавным ростом вероятности достижения порога события до единицы.

Рост вероятности перехода имеет резко выраженный ступенчатый характер. Это связано с тем, что коэффициент $b = \frac{\varepsilon - \xi}{\tau}$ в уравнении

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

равен 0, а само уравнение приобретает следующий вид:

$$\frac{d\rho(x,t)}{dt} = \frac{\varepsilon^2}{\tau}\frac{d^2\rho(x,t)}{dx^2} - \tau\frac{d^2\rho(x,t)}{dt^2}.$$

Упорядоченные переходы невозможны, так как исчезает член уравнения $\frac{d\rho(x,t)}{dx}$; остается член уравнения $\frac{d^2\rho(x,t)}{dx^2}$, который определяет случайное изменение (только случайные переходы). В этом случае член уравнения $\frac{d^2\rho(x,t)}{dt^2}$ определяет ускорение только случайных переходов, а упорядоченные переходы не ускоряются (нет члена уравнения вида $\frac{d\rho(x,t)}{dx}$).

Увеличение значения величин ε и ξ (при выполнении условия $\varepsilon > \xi$) изменяет величину плато (горизонтальный участок зависимости вероятности перехода через порог перколяции до второго участка резкого роста) на рисунке 3, однако общая зависимость вероятности перехода от времени качественно не изменяется.

6. ОСНОВНЫЕ СВОЙСТВА МОДЕЛИ ПРОГНОЗИРОВАНИЯ ИНФОРМАЦИОННЫХ СОБЫТИЙ НА ОСНОВЕ РЕШЕНИЯ КРАЕВОЙ ЗАДАЧИ ДЛЯ СИСТЕМ С РЕАЛИЗАЦИЕЙ ПАМЯТИ И САМООРГАНИЗАЦИЕЙ

1. Модель не основывается на статистических характеристиках процессов с заранее предполагаемым законом распределения и учитывает неопределенности в процессе возникновения событий в информационном пространстве.
2. Модель учитывает основные свойства реализации событий в информационном пространстве, такие как: неопределенность во времени их проявления, стохастичность, наличие памяти в системе в которой происходит событие, самоорганизация информации.
3. Модель показывает рост вероятности достижения порога прогнозируемого события в информационном пространстве практически сразу после начала процесса его развития, что связано с учетом памяти о предыдущих состояниях системы и возможности её самоорганизации.
4. Чем ближе величина начального состояния системы к величине порога реализации события, тем быстрее вероятность его достижения приближается к единице с течением времени.
5. Рост вероятности перехода через порог события имеет ступенчатый характер, а протяженность ступени во времени зависит от того насколько начальная величина состояния системы близка к порогу события.
6. Вероятность достижения порога события имеет осциллирующий характер. Причем чем дальше значение начальной величины начального состояния системы от порога события, тем сильнее проявляются осцилляции.
7. Модель позволяет оценить при заданной величине вероятности время реализации прогнозируемого новостного события.

СПИСОК ЛИТЕРАТУРЫ

1. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.
2. Pang-Ning Tan, Steinbach M., Kumar Vipin *Introduction to Data Mining*, Pearson Addison-Wesley, 2006.
3. Andrews N.O., Fox E.A., *Recent Developments in Document Clustering*, Department of Computer Science, Virginia Tech, Blacksburg, 2007.
4. Feldman R., Sanger J., *The Text Mining Handbok*, Cambridge: Cambridge University Press, 2009.
5. Lesko, S.A., Zhukov, D.O Trends, Self-Similarity, and Forecasting of News Events in the Information Domain, Its Structure and Director, *Proc. 2015 IEEE Int. Conf. on Smart City/SocialCom/SustainCom Together with DataCom 2015 and SC2*, 2015, pp. 870–873.
6. Zhukov D.O., Lesko S.A. Stochastic Self-Organisation of Poorly Structured Data and Memory Realisation in an Information Domain When Designing News Events Forecasting Models, *Proc. 2016 IEEE 14th Int. Conf. on Dependable, Autonomic and Secure Computing, 14th Int. Conf. on Pervasive Intelligence and Computing, 2nd Int. Conf. on Big Data Intelligence and Computing and Cyber Science and Technology Congress*, 2016, pp. 890–893.

The model of forecasting information events based on the solution of a boundary-value problems for systems with memory and self-organization

A.S. Sigov, D.O. Zhukov, T.Yu. Khvatova, E.G. Andrianova

One of the problems of forecasting news events is the development of models which allow you to work with a weakly structured information space of text documents. A distinctive feature of this news space is

the stochastic nature of the processes occurring in it, the presence of memory and the possibility of self-organization of information. It seems interesting to create a model for forecasting events which would be based on the stochastic dynamics of changes in the structure of news clusters (or information space states), and which would take into account the memory and self-organization of their structure. In this article describes the schemes of probabilities of transitions between states in the information space. On the basis of this description is derived a second-order nonlinear differential equation, and also formulated and solved a boundary value problem for forecasting news events. The analysis of the model described in the article shows the possibility of an increase in the probability of achieving the predicted event, almost immediately after the beginning of the process of changing the structure of news clusters, also the presence of sudden jumps and oscillations the value probability to reaching event.

Keywords: stochastic dynamics of changes in the states of the information system, self-organization, memory processes, news event threshold, information space, news cluster, news clustering.