УДК 004.942 DOI: 10.17587/prin.8.511-523

А. А. Артемов, инженер, соискатель, e-mail: artemsince2@yandex.ru, АНО Центр проблем стратегических ядерных сил Академии военных наук, г. Юбилейный, Московская обл.,
А. В. Галатенко, канд. физ.-мат. наук, ст. науч. сотр., e-mail: agalat@msu.ru,
Механико-математический факультет МГУ имени М. В. Ломоносова

Моделирование процесса эволюции содержания информационного пространства социума (Мем-грамм-модель)

Представлены модель и алгоритм для прогноза числа копий мема в будущем содержании информационного пространства социума (ИП). Глубина прогноза модели составляет одну календарную неделю на каждые три значения числа копий мема в предыдущие три недели. Модель является статистической и создана на основе анализа изменений содержания русскоязычного сегментам ИП с мая по август 2014 г., представленного коллекцией изменяющихся во времени корпусов текстов, сформированных из 300 000 наиболее цитируемых в социальных сетях русскоязычных публикаций различных интернет-СМИ. Приведены оценки качества модели, определяющие вероятность ошибки прогноза числа копий мема. Особенностью предложенного алгоритма является возможность его применения для прогноза числа копий мема для любого ИП, близкого по размеру и разнообразию русскоязычному сегменту ИП.

Ключевые слова: *мем-грамм-модель, языковая модель, биграмма, мем, информационное пространство, распределение Su-Джонсона, теорема о плотности распределения числа копий мемов, статистическая модель, алгоритм прогноза числа копий мемов*

Введение

На практике хорошо себя зарекомендовали два подхода к построению языковых моделей представления знаний (далее — языковая модель), основанных на анализе текстов на естественном языке с использованием методов машинного обучения. Первым из них является подход, основанный на анализе н-грамм — лемматизированной (в исходной форме) совокупности слов, полученных из текстов [1]. Вторым является подход на основе векторов, представляющих связь выбранного слова с контекстом (последовательностью нелемматизированных слов, рассматриваемых как уникальная языковая единица). Наиболее успешным примером векторной модели является модель word2vec [2]. В моделях знаний предлагается функция, способная предсказывать вероятность того, что определенная последовательность слов является допустимой (вероятной) для данного языка. Соответственно, в общем случае языковая модель представления знаний задается двойкой элементов $\langle C_T, \Pr(w_i)\rangle$, где C_T — корпус текстов T, представленных мультимножеством словосочетании w; Pr — функция вероятности, которая ставит в соответствие набору элементов $w_i = w_{n-1}w_n$ из мультимножества T число так, что $Pr : N[w_i] \to [0;1] | \forall w_i \in C_T,\ N[w_i] \in \mathbb{N}$. Базовая н-грамм-модель для биграмм представляется в следующем виде:

$$Pr\left(w_n|w_{n-1}\right)=\frac{N\left(\mathrm{w}_i=w_{n-1}\,w_n\right)}{\sum_k N\left(w_{n-1}w_n^k\right)}, \quad (1)$$

где $Pr\left(w_n|w_{n-1}\right)$ — вероятность встретить слово w_n после слова w_{n-1}; N — счетчик числа встречающейся последовательности слов w_{n-1} и w_n в исследуемых текстах.

Векторная модель word2vec представляется в виде

$$Pr\left(w_n|w_c\right)=\frac{\mathbf{e}^{u_{w_c}^T V_{w_i}}}{\sum_C \mathbf{e}^{u_{w_c}^T V_{w_i}}}, \quad (2)$$

где Pr — вероятность встретить слово w_i рядом (до или после) с выбранным набором слов w_c (именуемым контекстом), являющимся базисом $[w_c]$ для представления слова w_i вектором $\mathbf{w}_i$; $V_{\mathbf{w}_i}$ — булевский вес вектора $\mathbf{w}_i$ для выбранного слова w_i, который принимает значение 0 или 1 в зависимости от отсутствия или наличия в тексте избранного слова; $u_{w_c}^T$ — вес w_c в тексте T.

Представленные выше модели (1) и (2) и другие им подобные базируются на исходном мультимножестве словосочетаний (словаре), составленном из выбранных для обучения наборов текстов (корпусе). Точность модели зависит от мощности словаря и распределения частот словосочетаний в коллекции. Подход к выбору исходных текстов, как правило, экспертно-субъективный, эмпирический, без учета временного фактора. Отсюда получаем перечисленные далее недостатки этого подхода.

1. Исходные тексты могут изначально не представлять реальной ценности для решения задач в интересующей предметной области в текущее время, что означает "знания" неактуальны.

2. Исходные тексты могут иметь высокую частоту обновления содержащихся в них моделей знаний, а следовательно, через незначительное время собранные "знания" станут неактуальными. Необходимым условием работы с такой моделью является возможность определения эффективного периода актуальности для каждого объекта модели знания (н-граммы или слова). Достаточным условием является прогноз изменения частоты (или иной качественной характеристики) встреч заданного объекта модели знаний в текстах будущего (выбранного) периода актуальности.

Для решения задач, устраняющих недостаток п. 1 на практике, чаще всего прибегают к экспертной оценке и последующему "обогащению" данных на основе больших данных, наращивая мощность словаря (как правило, более 1 000 000 слов). Однако такой подход не позволяет решать задачи, направленные на устранение недостатков п. 2, где требуется, наоборот, сокращение объема данных, подлежащих анализу. Решения таких задач в существующих моделях пока не представлены. Как следствие, необходим подход, позволяющий учитывать оценку изменения языковой модели представления знаний $\langle C_T, Pr \rangle$ (сформированной на основе словосочетаний w корпуса текстов T) для прогноза вероятности использования словосочетания w в текстах будущего периода.

Решение поставленной задачи возможно двумя путями: "от частного к целому" — оценкой динамики изменения каждого объекта и последующей интегральной оценкой степени изменения (пригодности) модели в будущем; дедуктивно — оценкой динамики изменения модели в целом и последующим учетом этого изменения для каждого словосочетания в будущем. Основываясь на опыте работы [3], в которой рассмотрена модель для отслеживания изменения мемов, как устойчивых "мутирующих" словосочетаний (*meme tracker*, стэндфордская модель), и работе [4], где описана компьютерная модель процесса эволюции культуры (*memes and variations*, MAV), был сделан вывод о целесообразности выбора эволюционного подхода "от частного к общему" при решении задачи оценки изменения и прогноза содержания ИП. Исходя из данного выбора, были сформулированы две подзадачи для решения целевой задачи оценки изменения языковой модели:

— определить количественную меру отличия одной языковой модели вида $\langle C_T, Pr(w) \rangle$ относительно объектов (словосочетания) от другой модели такого же вида;

— разработать алгоритм для прогноза относительной частоты P словосочетания w_i в следующий момент времени (через выбранный интервал времени) для языка, представленного языковой моделью вида $\langle C_T, Pr(w) \rangle$.

Решением поставленных задач явилось создание мем-грамм-модели (МГМ). Описание решения первой задачи будет дано в другой статье. В настоящей статье представлено решение второй из перечисленных выше подзадач в виде описания модели и алгоритма ее использования для прогноза числа копий мема за недельный интервал времени.

1. Базовые понятия мем-грамм-модели

Далее представлена разработанная авторами модель на основе форматизированного понятия мема, представленного биграммой (мем-граммы) в следующий за настоящим моментом времени прогнозируемый временной период (далее кратко — следующий период).

В н-грамм-модели используются следующие подходы к представлению и оценке моделей знаний, сформированных на базе текста [2]:

— текст рассматривается как цепь Маркова (в том числе высших порядков);

— оценкой качества модели является мера правдоподобия, перплексия или иные критерии.

Предлагаемая модель базируется на данных подходах со следующими, указанными далее дополнениями.

1. Формализуется понятие ИП.

На входе модели поступают мультимножества последовательностей н-грамм w — текстов $\{T_1^\tau, T_2^\tau, \ldots, T_i^\tau\}$ с присвоенной им меткой времени τ, число текстов T_i^τ конечно. За начальное время τ_0 выбирается значение времени самого раннего текста, определяется значение времени самого позднего текста $\tau_{\min}$, затем выбирается дискретный интервал Δt. Предлагается два подхода к выбору Δt. Первый из них — выбрать наиболее "эффективный" интервал, такой, чтобы можно было обнаружить большинство изменений в языковых моделях при минимальном числе равных интервалов. Второй — "предикативный" подход, при котором выбираемый интервал Δt дает лучшие условия (группировки н-грамм) для решения задачи прогноза. Первый вариант рассмотрен в работе [5], второй будет представлен в данной статье (в описании процесса разработки Алгоритма 1). После определения шага дискретизации Δt рассчитывается число целых интервалов $u = \frac{(\tau_{\min} - \tau_0)}{\Delta t}$, $u \in \mathbb{N}$, задаются дискретные интервалы времени $\{T_1^{t_1}, T_2^{t_2}, \ldots, T_i^{t_p}\}$, $t_p = \tau_0 - p\Delta t$, $n \in \mathbb{N}$, $0 < n \leqslant u$, и формируется мультимножество текстов с дискретной меткой времени $\{T_1^{t_1}, T_2^{t_2}, \ldots, T_i^{t_p}\}$. Из данного мультимножества текстов формируется ИП.

Определение 1. Информационным пространством (ИП) с метрикой силы информационного воздействия на наследственность содержания ИП будем называть тройку $\left(\left\langle C^{Pr}, M^{\varphi}\right\rangle, \alpha^{IS}\right)$, где C^{Pr} — коллекция языковых (н-грамм) моделей $\left\langle C_p, Pr\left(w_k\right)\right\rangle$ с единой функцией оценки условной вероятности $Pr\left(w_j^p = w_{n-1}w_n\right)$ события получения после слова w_{n-1} н-граммы $w_j^p = w_{n-1}w_n$ в произвольном тексте $T_i^{t_p}$ момента времени p; M^{φ} — коллекция множеств M_p единиц наследственности — мемов m_i^p — общих элементов мультимножеств C_{p-k+1} и C_{p-k}, сформированных посредством процедуры φ; $p > 1$; $k, p, i, j \in \mathbb{N}$, $C_p = \left\{w_1^p, w_2^p, ..., w_n^p\right\}$ — мультимножество н-грамм текстов $T_n^{t_p}$ одного момента времени p; α^{IS}- метрика, определяющее меру различия двух языковых моделей C_p относительно всех мемов одного из множеств M_p (правого — при рассмотрении в перспективе, левого — при рассмотрении в ретроспективе).

Отметим, что в рамках данной работы рассматриваются языковые н-грамм-модели (для случая биграмм), в которых вероятность мема, представленного биграммой, определяется как относительная частота встречи биграммы, являющейся мемом, относительно всех биграмм ИП. Подробно подходам к мере α^{IS} будет посвящена другая статья.

2. На ИП задается мем-грамм-модель. С применением моделей на н-граммах решается задача предсказания вероятности того, что определенная последовательность слов является допустимой для данного языка C. В мем-грамм-модели решается иная задача — предсказать, что допустимая для языка C_p последовательность слов $w_{n-1}w_n$ в текущий момент времени p будет допустимой для языка C_{p+1} в следующий момент времени $p + 1$. Это условие означает — быть мемом $m_i^{p+1} = w_{n-1}w_n$ для языков C_p и C_{p+1}. Формализированная запись данного высказывания:

$$Pr\left(m_j^{p+1} = w_i^p\right) = \frac{N\left(w_j^{p+1} = m_i^{p+1}\right)}{\sum_q N\left(w_q^{p+1} = m_i^{p+1}\right)}, \qquad (3)$$

где m_i^{p+1} — мем будущего периода; $w_i^p = w_{n-1}w_n$ — известная биграмма языковой модели C_p из коллекций языковых моделей $C^{Pr} : w_i^p \in C_p \subseteq C^{Pr}$.

Таким образом, для оценки (3) необходимо построение алгоритма Λ для прогноза числа $N\left(w_j^{p+1} = m_i^{p+1}\right)$ копий словосочетания w_j^{p+1}, являющегося мемом m_j^{p+1}. Учитывая, что на практике интересен не любой алгоритм, а только те, которые обладают заданным параметром качества, для $\exists\Lambda$ необходимо потребовать оценку θ достоверности прогноза интервала числа копий мема $N^{\Lambda}\left(w_j^{p+1} = m_i^{p+1}\right) \in \left[N_a;\ N_b\right]$. Длина указанного интервала должна быть гарантированно меньше заданного ξ:

$$\begin{cases} P\left(N_a \leqslant N^{\Lambda}\left(w_j^{p+1} = m_i^{p+1}\right) \leqslant N_b\right) \geqslant \theta; \\ N_b - N_a \leqslant \xi. \end{cases} \qquad (4)$$

Определение 2. Мем-грамм-модель задана, если задано ИП $\left(\left\langle C^{Pr}, M^{\varphi}\right\rangle, \alpha^{IS}\right)$ и определен $\Lambda^{\theta,\xi}$ — алгоритм (или функция) прогноза в следующий момент времени $p + 1$ числа $N\left(w_j^{p+1} = m_i^{p+1}\right)$ экземпляров словосочетания w_j^{p+1}, являющегося мемом $m_j^{p+1} \in M_{p+1} \not\subset M^{\varphi}$, относительно словосочинений $w_q^p \in C_p \subset C^{Pr}$ ИП текущего момента времени p; $p > 1$, $p, q, i, j \in \mathbb{N}$, θ — оценка достоверности даваемого алгоритмом прогноза; ξ — максимальная ошибка даваемого алгоритмом прогноза. В соответствии с данными дополнениями необходимым условием для задания мем-грамм-модели является задание четверки элементов $\left(\left\langle C^{Pr}, M^{\varphi}\right\rangle, \alpha^{IS}, \Lambda^{\theta,\xi}\right)$.

На настоящее время представляется сложным разработать математический подход к определению функции $\Lambda^{\theta,\xi}$. Трудности на этом направлении в значительной степени обусловлены необходимостью корректного обоснования ряда гипотез, основанных на наиболее детальном изучении процессов изменения реального ИП. Для этих целей сначала был разработан инструментарий (программное обеспечение) для сбора и обработки реальных данных — 3000 публикаций русскоязычного сегмента Интернета за июнь—август 2014 г. [6], а затем на их основе были созданы языковые н-грамм-модели, коллекция которых определяет объект изучения — ИП русскоязычного сегмента Интернета (ИПР) [5]. Изучение процессов, происходящих в ИПР, позволило сформировать метрику силы информационного воздействия и подтвердить целесообразность использования концепции меметики для описания изменений ИП. Результаты такого изучения позволили сформировать гипотезу "Об эволюции популяции мемов", которая впоследствии подтвердилась в виде доказанной теоремы "О плотности вероятности функции числа копий мемов", что в итоге привело к возможности создания алгоритма $\Lambda^{\theta,\xi}$ для расчета прогнозного числа копий мема. Основой для подтверждения гипотезы, доказательства теоремы и определения алгоритма является разработанная и представленая в настоящей статье статистическая модель для прогноза числа копий мема в следующий за настоящим период времени.

2. Постановка задачи

Целью исследования изменений ИП является разработка статистической имитационной модели ИП для прогноза числа копий мема в следующий период времени. Такая модель необходима для проверки гипотезы "Об эволюции популяции мемов". В случае ее подтверждения необходимо разработать алгоритм для применения модели на практике.

Исходным объектом анализа являются объекты ИП $\left(\left\langle C^{Pr}, M^{\varphi}, \alpha^{IS}\right\rangle\right)$, представленного коллекцией

C^{Pr} мультимножеств C_p словосочетаний $w_i^p = w_{n-1}w_n$ (биграмм), полученных из множества текстов публикации T_i^τ произвольной длины с датировкой времени τ, которое можно разделить на "до" и "после" выбранного момента времени τ_0. Биграмма — словосочетание двух слов, приведенных к лемматизированной форме. Например, биграмма "мама мыть" получена из словосочетания "мама мыла". Соответственно, из множества текстов, представленных последовательностью биграмм, формируется коллекция мультимножеств C^{Pr}, а из общих биграмм мультимножеств C_{p-k+1}, $C_{p-k} \in C^{Pr}$ за последовательные моменты времени $p - k + 1$ и $p - k$ формируется коллекция мемов M^φ.

В дальнейшем будем использовать следующие обозначения:

$t_p = \begin{cases} \tau_0 - p\Delta t, & p < k \\ \tau_0 + p\Delta t, & p \geqslant k \end{cases}$ — момент времени, где $t_p \in \{t_1, t_2, ..., t_k = \tau_0, ..., t_n\}$ или кратко — момент времени p;

$N\left(w_i^p\right)$ — число копий w_i биграммы в момент времени p;

$Sum\left(N_p\right) = \sum_i N\left(w_i^p\right)$ — суммарное число копий биграмм в момент времени p;

$S_T^{IS} = \dfrac{\sum_p Sum\left(N_p\right)}{p}$ — среднее число биграмм за период времени T, используется для нормировки характеристик биграмм ИП;

$Sum\left(M_p\right) = \sum_i N\left(w_i^p = m_i^p\right)$ — суммарное число копий мемов в момент времени t_p, определяет размер ИП;

$NN\left(w_i^p\right) = \dfrac{N\left(w_i^p\right)}{S_T^{IS}}$ нормированное число копий w_i биграммы в момент времени p;

$NSum\left(N_p\right) = \dfrac{Sum\left(N_p\right)}{S_T^{IS}} \cdot 10^6$ — нормированный размер ИП, определяемый суммарным числом копий всех биграмм в момент времени t_p. Нормирование проводится по размеру ИП в один миллион биграмм (принято для удобства использования, поскольку число копий исследуемого ИП, принятого за базовое, измеряется в миллионах биграмм).

Перечислим задачи исследования.

1. Провести первичный анализ имеющихся данных: оценить количественное и качественное изменение содержания ИП во времени, выявить оптимальный период для анализа, провести исследование распределения числа копий биграмм — формирование отклика исследования с помощью логарифма от числа копий биграмм.

2. Провести статистическое исследование влияния характеристик биграмм и ИП на число копий биграмм в следующий дискретный момент времени $p + 1$.

3. Провести регрессионное моделирование распределения случайной величины числа копий в следующий дискретный момент времени $p + 1$: выявить наиболее значимые факторы, построить регрессионную модель для логарифма числа копий биграмм в следующий дискретный момент времени $p + 1$.

4. Подобрать теоретическое распределение для аппроксимации распределения остатков построенной регрессионной модели для логарифма числа копий биграмм в следующий дискретный момент времени $p + 1$: проверка гипотезы о принадлежности эмпирического распределения основным теоретическим законам.

5. Теоретическое исследование распределения числа копий биграмм в следующий дискретный момент времени $p + 1$ (вывод основной Теоремы).

6. Валидация построенной статистической модели распределения числа копий биграмм в следующий дискретный момент времени $p + 1$ на реальных данных.

7. Разработка алгоритма построения прогноза распределения числа копий биграмм в следующий момент времени $p + 1$ на основе исторических данных за последние три периода: $p - 2$, $p - 1$, p.

3. Статистический анализ данных.

В работе [5] автор детально описал, каким образом были выбраны источники данных и как проводилась их обработка для формирования ИП. Отметим кратко наиболее важную для дальнейшего изложения информацию.

3.1. Подготовка данных для анализа

В целях изучения изменения русскоязычного сегмента ИП было отобрано порядка 780 000 публикаций СМИ с 01.05.2014 по 01.04.2015, ранжируемых сервисом Meediametics.ru как наиболее цитируемые в социальных медиа. Публикации были обработаны методом, представленным в работе [5], в данные в виде массива строк: биграмма; число ее копий в указанную дату; дата (с шагом 1 день). Общее число биграмм (оригиналов по дням) в исходных данных составило 1 731 496.

Анализ исходных данных показал, что изменения в ИП имеют выраженную недельную сезонную составляющую. На рис. 1 приведены распределения по числу копий биграмм для каждого из дней недели. Как видно на графике, число копий биграмм постепенно нарастает с понедельника по пятницу и снижается в выходные дни.

Для снятия эффекта "влияния дня недели" целесообразно перейти от ежедневных данных к данным по неделям. При этом для построения модели будем использовать предысторию за два предыдущих периода.

Рис. 1. Гистограмма распределения числа копий биграмм по дням недели

В целях наглядного представления процессов изменения состава ИП приведем пример количественных характеристик десяти наиболее часто встречающихся мемов, представленных в табл. 1.

3.2. Группировка данных

Статистический анализ количественных характеристик биграмм показал, что имеет место высокая корреляционная зависимость числа копий мема $N\left(w_i^{p+1}\right)$ от динамики изменения в предыдущие три момента времени (табл. 2). Исходя из определения мема, для утверждения, что мем будет завтра, необходимо и достаточно появление хотя бы одной его копии. В соответствии с постановкой задачи, рассмотрим вероятность появления мема $P\left(N\left(w_i^{p+1}\right)\right)$ в зависимости от числа упоминаний этого мема в последовательные два предыдущих и текущий моменты времени. Тогда имеет место функциональная зависимость f так, что

$$P\left(N\left(w_i^{p+1}\right)\middle| N\left(w_i^{p-2}\right),\ N\left(w_i^{p-1}\right),\ N\left(w_i^{p}\right)\right) = \\ = f\left(N\left(w_i^{p-2}\right),\ N\left(w_i^{p-1}\right),\ N\left(w_i^{p}\right)\right).$$

На основе анализа исходных данных методом дерева решений были получены результаты, представленные

Таблица 1

Первые десять наиболее популярных мемов

№	Мем	Общее число упоминаний	Среднее число упоминаний в день
1	"об это"	41 261	156
2	"риа новость"*	36 270	137
3	"владимир путин"	30 932	117
4	"то число"	29 200	111
5	"кроме то"	22 808	86
6	"который быть"	19 461	73
7	"человек который"	17 336	65
8	"б б"	16 616	102
9	"самый дело"	16 034	61
10	"народный республика"	15 273	58

*Технический мем, вызванный ошибкой обработки публикации, связан с указанием источника в публикации, а не с ее содержанием.

Таблица 2

Группы мемов

Группа мемов	Число копий мема в момент времени			Вероятность появления мема в период $(p+1)$	Число записей* (объем выборки)
	$p-2$	$p-1$	p		
1	$N\left(w_i^{p-2}\right) \geqslant 1$	$N\left(w_i^{p-1}\right) \geqslant 2$	$N\left(w_i^{p}\right) \geqslant 4$	0,859	457 049
2	$N\left(w_i^{p-2}\right) \geqslant 1$	$N\left(w_i^{p-1}\right) \geqslant 2$	$N\left(w_i^{p}\right) < 4$	0,596	46 894
3	$N\left(w_i^{p-2}\right) \geqslant 1$	$N\left(w_i^{p-1}\right) < 2$	$N\left(w_i^{p}\right) \geqslant 4$	0,663	372 870
4	$N\left(w_i^{p-2}\right) \geqslant 1$	$N\left(w_i^{p-1}\right) < 2$	$N\left(w_i^{p}\right) < 4$	0,425	143 290
5	$N\left(w_i^{p-2}\right) = 0$	$N\left(w_i^{p-1}\right) \geqslant 2$	$N\left(w_i^{p}\right) \geqslant 4$	0,507	134 646
6	$N\left(w_i^{p-2}\right) = 0$	$N\left(w_i^{p-1}\right) \geqslant 2$	$N\left(w_i^{p}\right) < 4$	0,271	38 146
7	$N\left(w_i^{p-2}\right) = 0$	$N\left(w_i^{p-1}\right) < 2$	$N\left(w_i^{p}\right) \geqslant 4$	0,417	554 664
8	$N\left(w_i^{p-2}\right) = 0$	$N\left(w_i^{p-1}\right) < 2$	$N\left(w_i^{p}\right) < 4$	0,110	557 462

* Запись – уникальная комбинация идентификатора мема и номера недели.

в табл. 2. Данные результаты содержат оценки вероятности появления мема в зависимости от числа упоминаний за три предыдущих периода времени. Отметим, что под вероятностью появления мема понимается доля событий $N\left(w_i^{p+1}\right) > 0$ относительно суммы событий $N\left(w_i^{p+1}\right) > 0$ и $N\left(w_i^{p+1}\right) = 0$.

Таким образом, удалось выделить восемь групп исходных данных, отражающих различные предположения относительно $N\left(w_i^{p-2}\right)$, $N\left(w_i^{p-1}\right)$ и $N\left(w_i^{p}\right)$. Наиболее вероятное появление мема в следующем периоде получаем в группе 1 при условии, что $N\left(w_i^{p-2}\right) \geqslant 1$, $N\left(w_i^{p-2}\right) \geqslant 2$ и $N\left(w_i^{p}\right) \geqslant 4$, вероятность дальнейшего появления мема равна

$$P\left(N\left(w_i^{p+1}\right) > 0 \middle| N\left(w_i^{p-2}\right),\ N\left(w_i^{p-1}\right),\ N\left(w_i^{p}\right)\right) = 0{,}859.$$

Поскольку группа 1 — единственная группа, в которой у мема имеется полноценная история за последние три периода, то эту группу данных будем в дальнейшем использовать в качестве исходных данных для регрессионного моделирования. Выбор обусловлен тем, что мемы групп 5—8 не имеют историю за три периода, а это минимальное условие характеристики мема в поставленной задаче. Мемы групп 2—4 не имеют устойчивого роста, и если будут "успешны", то только когда перейдут в группу 1. Такой подход позволяет рационально использовать вычислительные и временные ресурсы для мемов с незначительным числом копий (группы 2—8), поскольку позволяется упростить задачу прогнозирования за счет снижения требования к точности прогноза: с "узкого" интервала числа копий н-граммы на "широкий" интервал — от одной и более копий н-граммы.

3.3. Регрессионное моделирование

Построение регрессионной модели будем проводить для совокупности данных, соответствующих группе 1 из табл. 2. Следует отметить, что для описания положительных случайных величин, аналогичных рассматриваемому в статье числу копий мема в момент времени (p + 1), чаще всего используют распределения из семейства экспоненциальных [7, 8]. В этом случае применяют мультипликативно-экспоненциальные регрессионные уравнения следующего вида:

$$y = \mathbf{e}^{\mathbf{BX}}\varepsilon, \qquad (5)$$

где y — отклик, $\mathbf{B}$ — вектор неизвестных параметров; $\mathbf{X}$ — вектор входных факторов; ε — случайная ошибка.

При логарифмировании левой и правой частей уравнения (5) получаем аддитивную множественную линейную регрессионную модель $\ln y = \mathbf{BX} + \ln\varepsilon$, параметры которой можно оценивать любым из классических методов, например, методом наименьших квадратов. В связи с этими соображениями в дальнейшем в качестве отклика регрессионной модели будем рассматривать величину $\ln\left(NN\left(w_i^{p+1}\right)\right)$. Для построения модели будем использовать шаговые алгоритмы регрессионного анализа [9]. В качестве регрессоров рассматривались функции числа копий мема в следующий период времени: от основных характеристик мема; от мультипликативных характеристик мема; от характеристик информационного пространства.

В ходе дальнейшего анализа исходные данные были разделены на обучающую и контрольную выборки. Обучающую выборку составили данные, полученные с 21-й по 30-ю недели, контрольную — данные, полученные с 31-й по 34-ю недели.

Все приведенные возможные регрессоры проверяли на корреляционную зависимость с откликом. Для этого применяли пакет прикладных программ для статистического анализа JMP SAS: проводили расчет коэффициентов корреляций и их значимости. В качестве корреляций использовали коэффициенты Пирсона и Спирмена. В результате проверки было выявлено, что все факторы (регрессоры), в той или иной мере характеризующие размерность ИП, не оказывают статистически значимого влияния на $\ln\left(NN\left(w_i^{p+1}\right)\right)$. Статистически значимые коэффициенты корреляции по обучающей и кон-

Таблица 3

Топ 5 регрессоров, оказывающих статистически значимое влияние на отклик

Возможный регрессор	Коэффициент корреляции с $\ln\left(NN\left(w_i^{p+1}\right)\right)$	
	Обучающая выборка	Контрольная выборка
$\ln\left(NN\left(w_i^{p}\right)\right)$	0,6692	0,6641
$\ln\left(NN\left(w_i^{p-1}\right)\right)$	0,6209	0,6248
$\ln\left(NN\left(w_i^{p-2}\right)\right)$	0,6116	0,5898
$NN\left(w_i^{p}\right) NN\left(w_i^{p-1}\right)$	0,1485	0,1605
$NN\left(w_i^{p-1}\right) NN\left(w_i^{p-2}\right)$	0,1434	0,1460

Таблица 4

Показатели качества модели

Показатель	Обучающая выборка	Контрольная выборка
R^2	0,527452	0,516364
Средняя квадратическая ошибка	0,639258	0,648981
Число наблюдений	292 307	100 181

Таблица 5

Статистические характеристики оценок параметров

Регрессор	Оценки параметров*	Стандартная ошибка	t-статистика	P-значения	95 %-ный доверительный интервал	
					Нижняя граница	Верхняя граница
a_0	−0,069779	0,003311	−21,07	<0,0001	−0,07627	−0,06329
$\ln\left(NN\left(w_i^p\right)\right)$	0,493715	0,002325	212,35	<0,0001	0,489158	0,498272
$\ln\left(NN\left(w_i^{p-1}\right)\right)$	0,2133316	0,002065	103,31	<0,0001	0,209284	0,217379
$\ln\left(NN\left(w_i^{p-2}\right)\right)$	0,2366009	0,00179	132,2	<0,0001	0,233093	0,240109

* Оценки получены на основе метода наименьших квадратов.

трольной выборкам были получены только для десяти из двенадцати рассмотренных регрессоров. Пять из наиболее значимых регрессоров приведенны в табл. 3.

Количество моделей, которые теоретически могли бы быть построены с использованием даже 12 регрессоров, очень велико и составляет более чем $\sum_{k=1}^{12} C_{12}^k = 2^k - 1 = 4095$ единиц. Задача выбора наилучшей модели решалась с использованием специальных итерационных шаговых алгоритмов [9], реализованных в большинстве современных статистических программных пакетов. Из всего множества рассмотренных моделей (с различными регрессорами) наиболее предпочтительной (по значению R^2) оказалась авторегрессионная модель с лаговыми переменными (значение которой взято за предыдущие моменты времени) от периодов времени p, $(p - 1)$ и $(p - 2)$ следующего вида:

$$\ln\left(NN\left(w_i^{p+1}\right)\right) = a_0 + a_1 \ln\left(NN\left(w_i^p\right)\right) + \\ + a_2 \ln\left(NN\left(w_i^{p-1}\right)\right) + a_3 \ln\left(NN\left(w_i^{p-2}\right)\right) + \varepsilon. \quad (6)$$

Таблица 6

Перцентильная группировка наблюдений по значению регрессии

Перцентили	Объем подвыборки		Среднее значение регрессии	
	Обучения	Контроль	Обучения	Контроль
0,005	2024	768	0,62828	0,62800
0,025	5807	1927	0,72407	0,72539
0,05	6955	2293	0,79669	0,79507
0,1	14 162	4844	0,87986	0,87881
0,2	29 517	9881	0,99956	0,99985
0,3	29 174	9849	1,13747	1,13615
0,4	29 343	10 048	1,26412	1,26446
0,5	29 211	10 091	1,39745	1,39664
0,6	29 192	9816	1,54322	1,54306
0,7	29 229	10 560	1,71909	1,71600
0,8	29 232	10 237	1,94535	1,94405
0,9	29 230	9761	2,28319	2,28281
0,95	14 616	4942	2,70246	2,69844
0,975	7308	2448	3,11692	3,12248
0,995	5846	2207	3,69460	3,70338
1	1461	509	4,71040	4,65646

Результаты оценивания параметров модели (6) и оценка ее качества на обучающей и контрольной выборках приведены в табл. 4 и 5. Как видно из данных табл. 4, на обучающей и контрольной выборках с помощью модели удается охватить более 50 % вариаций отклика (значения $R^2 > 0{,}5$).

Отметим, что все параметры модели обладают хорошей значимостью (табл. 6).

С использованием данной модели прогнозируемая величина $\ln\left(Sn_i(t+1)\right)$ может быть получена по следующей формуле:

$$\ln\left(NN\left(w_i^{p+1}\right)\right)_{\text{Прогноз}} = -0{,}069779 + 0{,}493715 \times \\ \times \ln\left(NN\left(w_i^p\right)\right) + 0{,}2133316 \ln\left(NN\left(w_i^{p-1}\right)\right) + \\ + 0{,}2366009 \ln\left(NN\left(w_i^{p-2}\right)\right). \quad (7)$$

На рис. 2 представлены графики рассеивания наблюдаемых значений переменной $\ln\left(NN\left(w_i^{p+1}\right)\right)$

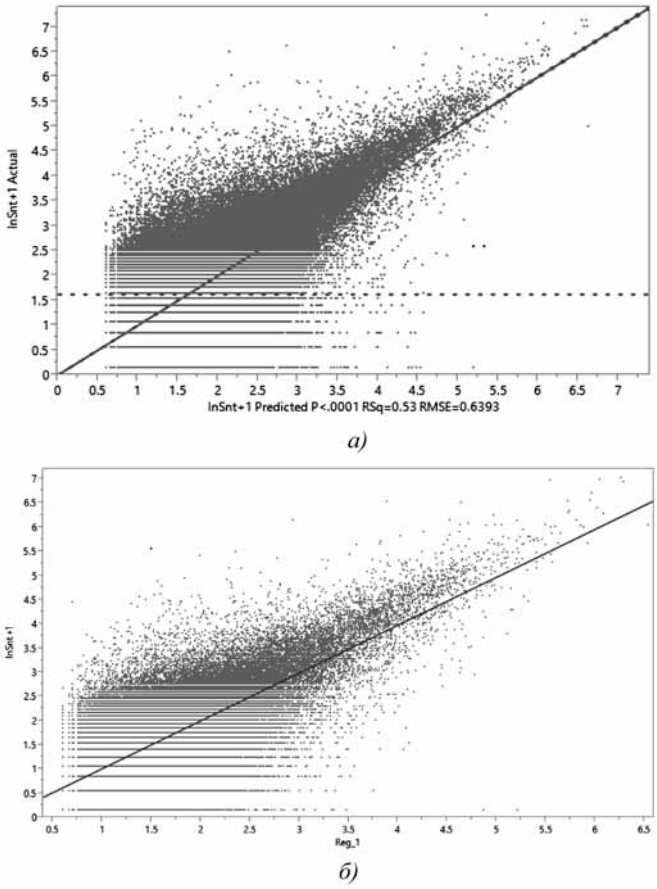

Рис. 2. Диаграмма рассеивания наблюдаемых значений в зависимости от прогнозированных значений для переменной $\ln\left(NN\left(w_i^{p+1}\right)\right)$:
а — по обучающей выборке; *б* — по контрольной выборке

в зависимости от значений прогнозируемых величин соответственно для обучающей и контрольной выборок. Как видно на рис. 2, в обоих случаях наблюдается ярко выраженная неоднородность исходных данных, а именно, при уменьшении значений переменных вариация данных существенно возрастает.

Оценки параметров, представленных ранее в табл. 4 и 5, были получены с использованием обычного метода наименьших квадратов [9], который не учитывает возможную неоднородность исходных данных. Для корректного оценивания параметров необходимо получить более детальную информацию о распределении отклика. Для решения этой задачи была проведена группировка значений регрессии по децентилям согласно табл. 6.

Такая группировка дает возможность сформировать однородные группы, в которых наблюдается одинаковое распределение отклика, независящее от регрессоров. В каждой такой группе корректно провести процедуру оценки параметров уже не для остатков (разницы фактического и прогнозируемого значений), а для исходных значений отклика. Таким способом устраняется влияние значения регрессоров.

3.4. Подбор распределения для отклика

Для подбора подходящего распределения случайной величины $\ln\left(NN\left(w_i^{p+1}\right)\right)$ выделим сегмент, имеющий наиболее ярко выраженную взаимосвязь межу регрессией (7) и $\ln\left(NN\left(w_i^{p+1}\right)\right)$. Этот сегмент соответствует значению перцентиля от 0,995 до 1 в табл. 6, правой верхней части облака рассеивания точек на рис. 2. Процесс оценивания формы распределения будем проводить по остаткам регрессии для сегмента перцентиля от 0,995 до 1, используя обучающую и контрольную выборки. Целесообразность такого выбора объясняется тем обстоятельством, что на практике представляет интерес выявления мема следующего периода не на самых ранних стадиях зарождения (2—3 копии), когда вероятность ложного детектирования весьма велика, а на стадии начального роста (10...20 копий), где вероятность ошибки меньше. Такой стадии и соответствует сегмент перцентиля от 0,995 до 1. Цель данного анализа — подобрать наилучшее распределение для случайной величины. Зная закон распределения остатков, будем знать закон распределения для отклика (так как они будут совпадать). В качестве возможных распределений будем рассматривать наиболее известные распределения, используемые в математической статистике для описания распределений регрессионных остатков: нормальное, мультинормальное, Лапласа, Коши, логистическое, минимальных значений, L-распределение и распределение Su-Джонсона [8, 9].

Для каждого из перечисленных распределений была проведена проверка статистической гипотезы о принадлежности выборки указанному закону распределения, т. е. проверки того, что эмпирическое распределение соответствует предполагаемой модели. Проверка осуществлялась с использованием критерия согласия Колмогорова [8, 9].

В табл. 7 приведены результаты проверки согласия имеющихся данных по остаткам модели с перечисленными выше распределениями, а также полученные по выборке параметры функций распределений. Для дальнейшей работы было выбрано распределение Su-Джонсона, так как ему соответствует минимальное (0,58) значение статистики Колмогорова. Таким образом, на основании критерия согласия Колмогорова была принята гипотеза, что плотности распределения логарифма числа копий мемов в ИП описываются распределением Su-Джонсона.

Таблица 7

Результат проверки согласия

Распределение	Статистика Колмогорова	Параметры распределения*			
		Масштаб	Сдвиг	Асимметрия	Эксцесс
Лапласа	2,28	0,3251	0,2524	—	—
Нормальное	6,54	0,5979	0,1755	—	—
Коши	2,64	0,1944	0,2614	—	—
Логистическое	2,38	0,2272	0,242	—	—
Минимальных значений	3,51	0,4126	0,4014	—	—
L-распределение	1,55	0,1819	0,5701	0,5164	1,5936
Su-Джонсона	0,58	0,261	0,4047	0,4618	0,9661

* Оценку проводили на основе метода минимального расстояния "Омега-большое квадрат Мизеса". Данный метод обладает большей устойчивостью, нежели классический метод максимального правдоподобия [8, 9].
Примечание. Прочерк "—" означает недопустимость данных параметров для соответствующих распределений.

4. Модель на основе распределения Su-Джонсона

Применение распределения Su-Джонсона для моделирования процесса изменения числа копии мема невозможно без решения двух задач. Первая задача — разработать и доказать корректность преобразования плотности распределения Su-Джонсона логарифма случайной величины в плотность распределения случайной величины. Вторая задача — определить зависимость изменения дополнительных параметров распределения Su-Джонсона — масштаба, эксцесса, асимметрии и сдвига от логарифма числа копий.

Представим краткое описание предлагаемых подходов для решения этих задач и полученных результатов их применения.

4.1. Распределение Su-Джонсона применительно к результатам статистического анализа

Обобщим полученные результаты в виде теоретических выкладок. Для этого рассмотрим случайную величину ξ, распределенную по закону Su-Джонсона с плотностью распределения [8, 9]

$$f(x,\gamma,\delta,\lambda,\mu)=\frac{\delta}{\lambda\sqrt{2\pi}}\frac{1}{\sqrt{\left(\frac{x-\mu}{\lambda}\right)^2+1}}\times$$
$$\times\exp\left\{-\frac{1}{2}\left[\gamma+\delta\ln\left(\frac{x-\mu}{\lambda}+\sqrt{\left(\frac{x-\mu}{\lambda}\right)^2+1}\right)\right]^2\right\}, \quad (8)$$

где $\delta > 0$ — параметр эксцесса; γ — параметр асимметрии; $\lambda > 0$ — параметр масштаба; μ — параметр сдвига; $y\in(-\infty;+\infty)$.

Основные характеристики распределения (4):

Математическое ожидание

$$E\xi=\mu-\lambda\exp\left(\frac{\delta^{-2}}{2}\right)\sinh\left(\frac{\gamma}{\delta}\right).$$

Дисперсия

$$D\xi=\frac{\lambda^2}{2}\left(\mathbf{e}^{\delta^{-2}}-1\right)\left(\mathbf{e}^{\delta^{-2}}\cosh\left(\frac{2\gamma}{\delta}\right)+1\right). \quad (9)$$

Теорема 1. О плотности распределения композиции функции распределения Su-Джонсона и экспоненциальной функции (о плотности распределения числа копий мемов).

Пусть величина ξ имеет плотность распределения $f_\xi(x,\delta,\gamma)$ (8), тогда случайная величина $\eta=\mathbf{e}^\xi$ имеет плотность распределения

$$g(x,\gamma,\delta,\lambda,\mu)=\frac{1}{x}\frac{\delta}{\lambda\sqrt{2\pi}}\frac{1}{\sqrt{\left(\frac{\ln x-\mu}{\lambda}\right)^2+1}}\times$$
$$\times\exp\left\{-\frac{1}{2}\left[\gamma+\delta\ln\left(\frac{\ln x-\mu}{\lambda}+\sqrt{\left(\frac{\ln x-\mu}{\lambda}\right)^2+1}\right)\right]^2\right\}, \quad (10)$$

где $\delta > 0$ — параметр эксцесса; γ — параметр асимметрии; $\lambda > 0$ — параметр масштаба; μ — параметр сдвига; $y\in(-\infty;+\infty)$.

Доказательство. Сформулируем в виде вспомогательной теоремы известное из теории вероятностей утверждение о преобразовании плотности функции распределения с помощью монотонной функции [8]: пусть ξ имеет плотность распределения $f_\xi(x)$ (8) и функция $h(x)$ — монотонна, тогда случайная величина $\eta=\eta(\xi)$ имеет плотность распределения $f_\eta(x)=\left(h^{-1}(x)\right)'f_\xi\left(h^{-1}(x)\right)$.

Функция $h(x) = \mathbf{e}^x$ монотонна на всем промежутке значений x — условия вспомогательной теоремы удовлетворены, обратная функция для $h(x) = \mathbf{e}^x$: $h^{-1}(x) = \ln(x)$. Производная от $\left(h^{-1}(x)\right)' = \dfrac{1}{x}$. Для формулы (8) с учетом вспомогательной теоремы получаем исходное распределение (10).

Смысл Теоремы 1 раскрывается в возможность перехода от функции логарифма числа копий мема к функции от числа копий мема: получения оценки параметров распределения случайной величины $\ln\left(NN\left(w_i^{p+1}\right)\right)$ (8) для построения доверительных интервалов распределения случайной величины $NN\left(w_i^{p+1}\right)$ (10).

На рис. 3 представлены частные случаи распределений случайных величин ξ (8) и $\eta = \mathbf{e}^{\xi}$ (10) при нулевом сдвиге и единичном масштабе ($\mu = 0$; $\lambda = 1$).

4.2. Исследование параметров распределения Su-Джонсона

Изучим влияние значений регрессии на параметры распределения. Для этой цели проведем оценивание параметров распределения Su-Джонсона для отклика $\ln\left(NN\left(w_i^{p+1}\right)\right)$, построенных для каждой перцентильной группы от 0,005 до 1 на основе регрессионной модели (7). Оценку проводили на основе метода минимального расстояния "Омега-большое квадрат Мизеса" с применением программы для статистического анализа данных ISW 4.4.1.98 профессора Б. Ю. Лемешко[1].

В частности, для описания параметра асимметрии ***A*** подходят две модели: первые десять групп с коэффициентом детерминации (R^2) 0,9885 описываются логарифмической кривой, а остальные группы с коэффициентом детерминации 0,9725 — линейной функцией

$$A = \begin{cases} 4,368\ln(x) - 0,8087, & x < 1,719 \\ -0,4685x + 2,2014, & x \geqslant 1,719. \end{cases} \qquad (11)$$

Для описания параметра эксцесса ***E*** подходят две модели: первые десять групп с коэффициентом детерминации 0,99325 описываются квадратичной параболой, а остальные группы с коэффициентом детерминации 0,9947 — линейной функцией

$$E = \begin{cases} 3,0039x^2 - 9,3989x + 9,9661, & x < 1,719 \\ -0,5393x + 3,5559, & x \geqslant 1,719. \end{cases} \qquad (12)$$

Для описания параметра масштаба ***M*** можно использовать одну модель: с коэффициентом детерминации 0,999 эта зависимость логарифмической кривой

$$M = -1,112\ \ln(x) + 2,1283. \qquad (13)$$

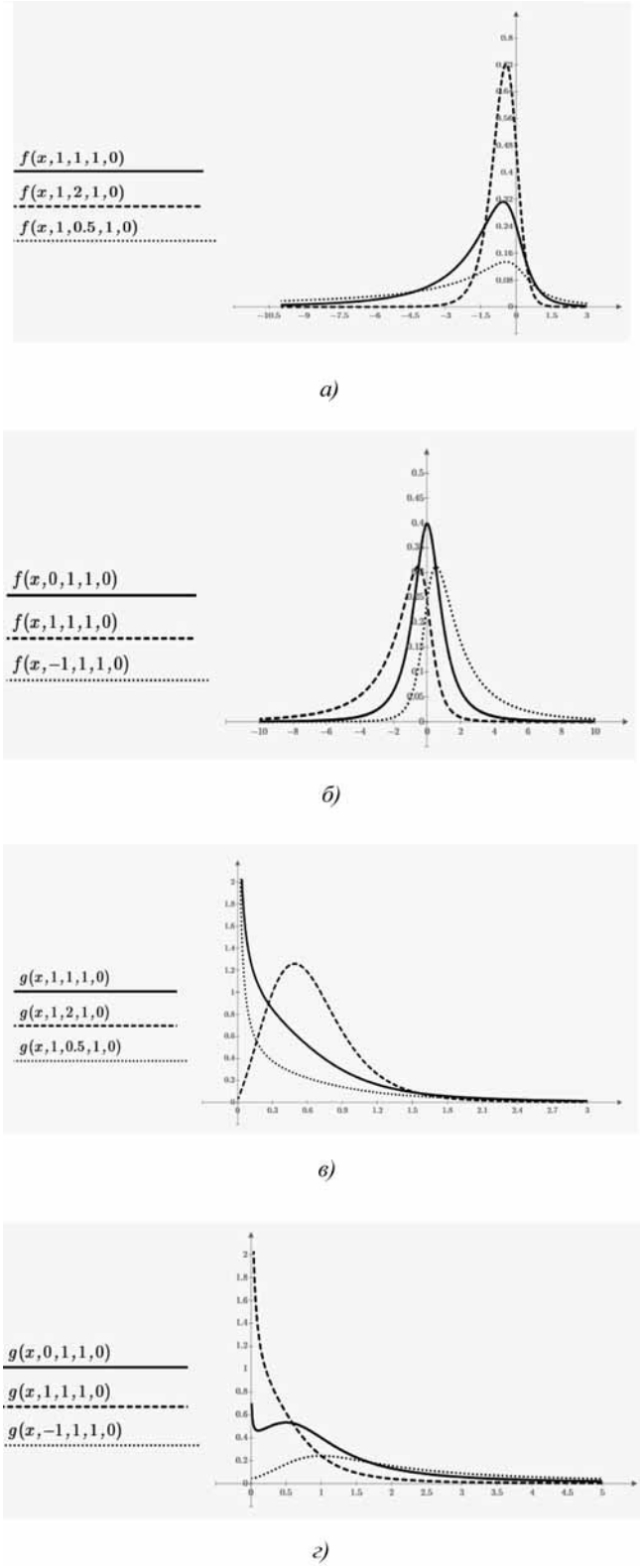

Рис. 3. Графики изменения плотности распределения Su-Джонсона в зависимости от эксцесса δ (*а*); асимметрии γ (*б*); масштаба (*в*) и сдвига λ (*г*)

[1] Программа доступна на сайте https://www.ami.nstu.ru/~headrd/

Для описания параметра сдвига $\boldsymbol{S}$ подходят две модели: первые десять групп с коэффициентом детерминации 0,9874 описываются первой линейной моделью, а остальные группы с коэффициентом детерминации 0,9947 — второй линейной моделью

$$S = \begin{cases} 3,1139x - 2,6093, & x < 1,719 \\ 0,7942x + 1,1385, & x \geqslant 1,719. \end{cases} \quad (14)$$

4.3. Валидация построенной регрессионной модели

Для проверки корректности построенной модели (7) было проведено ее тестирование на обучающей (число записей 292 307) и контрольной (число записей 100 181) выборках при условии, что все параметры модели (11)—(14) были оценены на основе обучающей выборки. В качестве оценки адекватности модели будем рассматривать качество построенных доверительных интервалов. В табл. 8 и 9 для сравнения приведены результаты тестирования при аппроксимации нормальным распределением и распределением Su-Джонсона.

Как видно из данных табл. 8, 9, средний модуль ошибки на обучающей и контрольной выборке при аппроксимации нормальным распределением составляет соответственно 1,9 и 1,6 %. При аппроксимации распределением Su-Джонсона средний модуль ошибки на обучающей и контрольной выборках существенно ниже и составляет соответственно 1,0 и 0,8 %.

5. Алгоритм прогноза числа копий мема

Прогноз количества копий мема в следующий момент времени для произвольного ИП, удовлетворяющий закон сохранения разнообразия, может быть осуществлен с использованием разработанной имитационной модели эволюции ИП по алгоритму, приведенному ниже.

Алгоритм прогноза числа копий мема в ИП следующего момента времени

1. Вход: S_T^{IS}, $N\left(w_i^p\right)$, $N\left(w_i^{p-1}\right)$, $N\left(w_i^{p-2}\right)$, $Sum\left(N_p\right)$ // размер эталонного ИП, число копий мема за три интервала времени, размер реального ИП.

1.1. $k = S_T^{IS} \cdot 10^{-6}$ // определение нормировочного коэффициента.

1.2. $NN\left(w_i^p\right) = \dfrac{N\left(w_i^p\right)}{k}$, $NN\left(w_i^{p-1}\right) = \dfrac{N\left(w_i^{p-1}\right)}{k}$, $NN\left(w_i^{p-2}\right) = \dfrac{N\left(w_i^{p-2}\right)}{k}$ // нормировка числа копий мема.

Таблица 8

Процент попадания значения $\ln\left(NN\left(w_i^{p+1}\right)\right)$ в доверительный интервал при аппроксимации нормальным распределением

Выборка	Доверительный интервал, %						
	95,0	90,0	75,0	50,0	25,0	10,0	5,0
Обучающая	93,0	87,9	75,5	52,8	23,6	7,5	3,4
Контрольная	92,6	87,1	74,4	52,3	24,7	8,3	3,9
Ошибочное попадание в доверительный интервал, %							
Обучающая	−2,0	−2,1	0,5	2,8	−1,4	−2,5	−1,6
Контрольная	−2,4	−2,9	−0,6	2,3	−0,3	−1,7	−1,1

Таблица 9

Процент попадания значения $\ln\left(NN\left(w_i^{p+1}\right)\right)$ в доверительный интервал при аппроксимации распределением Su-Джонсона

Выборка	Доверительный интервал, %						
	95,0	90,0	75,0	50,0	25,0	10,0	5,0
Обучающая	96,4	91,3	76,4	51,8	25,3	9,4	4,6
Контрольная	96,0	90,5	75,3	51,6	26,5	10,5	5,3
Ошибочное попадания в доверительный интервал, %							
Обучающая	−2,0	−2,1	0,5	1,8	0,3	−0,6	−0,4
Контрольная	1,0	0,5	0,3	1,6	1,5	0,5	0,3

2. Case: $\left(\left\{NN\left(w_i^p\right), \left(w_i^{p-1}\right), \left(w_i^{p-2}\right)\right\} \Rightarrow GroupID\left(w_i^{p+1}\right) = [1...8]\right)$ // определение группы мемов.

3. If $GroupID\left(w_i^{p+1}\right) \neq 1$ **Then**.

4. Выход 1: $N\left(w_i^{p+1}\right) = 1$; $P\left(w_i^{p+1}\right) = P(GroupID)$ определяется вероятность, что будет хоть одна копия мема в следующий интервал времени.

5. Else

5.1. $x = -0,069779 + 0,493715\ln\left(NN\left(w_i^p\right)\right) +$

$+ 0,2133316\ln\left(NN\left(w_i^{p-1}\right)\right) + 0,2366009\ln\left(NN\left(w_i^{p-1}\right)\right)$ // прогноз логарифма числа копий.

5.2. $\gamma = \begin{cases} 3,1139x - 2,6093, & x < 1,719 \\ 0,7942x + 1,1385, & x \geqslant 1,719. \end{cases}$

5.3. $\delta = \begin{cases} 4,368\ln(x) - 0,8087, & x < 1,719 \\ -0,4685x + 2,2014, & x \geqslant 1,719. \end{cases}$

5.4. $\lambda = \begin{cases} 3,0039x^2 - 9,3989x + 9,9661, & x < 1,719 \\ -0,5393x + 3,5559, & x \geqslant 1,719. \end{cases}$

5.5. $\mu = -1,112\ln(x) + 2,1283.$

5.6. $$x1 = \int_0^{1000} \frac{1}{x}\frac{\delta}{\lambda\sqrt{2\pi}}\frac{1}{\sqrt{\left(\frac{\ln x - \mu}{\lambda}\right)^2 + 1}} \times$$

$$\times \exp\left(-\frac{1}{2}\left(\gamma + \delta\ln\left(\frac{\ln x - \mu}{\lambda} + \sqrt{\left(\frac{\ln x - \mu}{\lambda}\right)^2 + 1}\right)\right)^2\right) \times$$

$$\times\, dx = 0,05.$$

5.7. $$x2 = \int_0^{1000} \frac{1}{x}\frac{\delta}{\lambda\sqrt{2\pi}}\frac{1}{\sqrt{\left(\frac{\ln x - \mu}{\lambda}\right)^2 + 1}} \times$$

$$\times \exp\left(-\frac{1}{2}\left(\gamma + \delta\ln\left(\frac{\ln x - \mu}{\lambda} + \sqrt{\left(\frac{\ln x - \mu}{\lambda}\right)^2 + 1}\right)\right)^2\right) \times$$

$$\times\, dx = 0,05.$$

5.8. **Выход 2:** $P\left(x1k \leqslant N\left(w_i^{p+1}\right) \leqslant x2k\right) = 0,9$ // определение интервала числа копий мема с 90 %-ной вероятностью.

6. **End**.

Отметим, что результат работы алгоритма — это возможность получения прогноза числа копий мема в диапазоне единиц и десятков единиц. Качество данного результата можно считать хорошим, поскольку возможно детектировать мемы на ранних этапах их зарождения. Наглядное сравнение данной оценки можно провести по количественным показателям "зрелых" мемов (см. табл. 1), насчитывающих от сотен до тысяч копий в неделю.

Заключение

Полученные в данном исследовании результаты кратко можно сформулировать в виде следующих положений.

- На основе анализа динамики изменения числа копий 200 000 биграмм, являющихся мемами (обучающая выборка), была разработана математическая модель, решающая задачу прогноза числа копий мема в ИП в следующий недельный интервал времени в зависимости от изменения числа копий данного мема за три предыдущих недельных интервала.
- Сформулирована и доказана теорема о распределении числа копий мема в следующий дискретный период времени $p + 1$.
- Валидация построенной статистической модели распределения числа копий мема в следующий дискретный момент времени на реальных данных показала высокую точность попадания реальных значений отклика в заданные 95 %-ные доверительные интервалы со средней погрешностью не более 1 %. Оценка качества разработанного модели (алгоритма), получена в результате проведения 100 000 экспериментов по прогнозу числа копий мема из контрольной выборки.

Авторы выражают благодарность научным руководителям д-ру физ.-мат. наук, проф. МГУ имени М. В. Ломоносова Валерию Александровичу Васенину и д-ру техн. наук, проф. Сергею Павловичу Расторгуеву за помощь в разработке мем-грамм-модели.

Список литературы

1. **Ramisch C.** N-gram models for language detection. URL: http://ru.scribd.com/doc/244210124/N-gram-Models-for-Language-Detection#
2. **Mikolov T., Chen K., Corrado G., Dean J.** Efficient estimation of world representation in vector space. URL: http://arxiv.org/pdf/1301.3781.pdf
3. **Leskovec J., Backstrom L., Kleinberg J.** Meme-tracking and the dynamics of the news cycle //Proceedings of the 15th ACM SIGKDD — International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2009. P. 497—506.
4. **Gabora L.** Meme and variations: A computational model of cultural evolution // 1993 Lectures in complex systems. Addison Wesley, 1995. P. 471—485.
5. **Артёмов А. А.** Эксперимент по моделированию процесса эволюции содержания информационного пространства социума (с применением мем-грамм-модели) // Программная инженерия. 2016. Т. 7, № 7. С. 291—306.
6. **Артёмов А. А.** Модель оценки уровня угроз информационных вызовов плану содержания информационного пространства социально-телекоммуникационной системы // Информационные войны. 2015. № 3 С. 83—97.
7. **Шуленин В. П.** Математическая статистика. Ч. 1. Параметрическая статистика: учеб. Томск: Изд-во НТЛ, 2012. 540 с.
8. **Кендалл М., Стьюарт А.** Статистические выводы и связи. М.: Наука, 1973.
9. **Айвазян С. А., Енюков И. С., Мешалкин Л. Д.** Прикладная статистика. М.: Финансы и статистика, 1985. 488 с.

Simulation of Society Information Space Evolution Process (The Meme-Gram-Model)

A. A. Artemov, artemsince2@ya.ru, Center of Strategic Nuclear Forces Research at Academy of Military Sciences, Moscow Region, Yubileiny, 141090, Russian Federation, **A. V. Galatenko,** agalat@msu.ru, Mechanics and Mathematics Faculty of Lomonosov Moscow State University, Moscow, 119234, Russian Federation

Corresponding author:

Artemov Artem A., Research Associate, Center of Strategic Nuclear Forces Research at Academy of Military Sciences, Moscow Region, Yubileiny, 141090, Russian Federation,
E-mail: artemsince2@ya.ru

Received on September 04, 2017
Accepted on September 18, 2017

The article presents a model and an algorithm for predicting the number of copies of a meme in the future content of the society's information space (IP). Authors presents a hypothesis that the information space of contemporary society can be described by the dynamics of changing of informational messages, which are spread by the means of mass communication technologies. The prediction period of the model is one calendar week for every three values of number of meme copies in three previous weeks. The model is statistical and was made based on the analysis of changes in the content of Russian-speaking segment of information space from May to August 2014 presented by a collection of time-varying text corpuses which were generated from 300 000 most cited in social media publications of various Internet media. In the article authors evaluate the quality of the model, which show the probability of errors in the forecast of number of copies of a meme. A special feature of the algorithm is the possibility of its application for prediction of number of meme copies for any IP with similar size and diversity parameters to the Russian segment of IP. The results of the study can be summarized in the following statements: 1) based on the analysis of the changes dynamics in the number of 200 000 bigrams copies or memes (training sample), we have developed a mathematical model that solves the problem of predicting the number of meme copies in IP for the next weekly interval depending on the change in the number of copies of the meme in the 3 previous weekly intervals; 2) a theorem on the distribution of the number of meme copies in the next discrete period of time p + 1 has been produced and proved; 3) validation of the constructed statistical model of the number of meme copies distribution in the next discrete moment of time on real data showed a high accuracy of the actual values of the response within the given 95 % confidence intervals with an average error of less than 1 %. Quality assessment of the developed model (algorithm) has been obtained as a result of conducting 100,000 experiments to forecast the number of copies of a meme in the control sample (validation set).

Keywords: *Mem-Gram-Model, language model, n-gram, bigram, memetic, meme, information space, Su-Johnson's distribution, distribution law, number of memes' copies, statistical model, algorithm.*

For citation:

Artemov A. A., Galatenko A. V. Simulation of Society Information Space Evolution Process (The Meme-Gram-Model), *Programmnaya Ingeneria*, 2017, vol. 8, no. 11, pp. 511—523.

DOI: 10.17587/prin.8.511-523

References

1. **Ramisch C.** N-gram models for language detection, available at: http://en.scribd.com/doc/244210124/N-gram-Models-for-Language-Detection#.
2. **Mikolov T., Chen K., Corrado G., Dean J.** Efficient estimation of the world representation in vector space, available at: http://arxiv.org/pdf/1301.3781.pdf
3. **Leskovec J., Backstrom L., Kleinberg J.** Meme-tracking and the dynamics of the news cycle, *Proceedings of the 15th ACM SIGKDD — International Conference on Knowledge Discovery and Data Mining,* New York, ACM, 2009, pp. 497—506.
4. **Gabora L.** Meme and variations: A computational model of cultural evolution, *1993 Lectures in complex systems*, Addison Wesley, 1995, pp. 471—485.
5. **Artemov A. A.** Eksperiment po modelirovaniju processa jevoljucii soderzhanija informacionnogo prostranstva sociuma (s primeneniem mem-gramm-modeli) (An experiment of simulation the process of society information space content evolution (using the meme-gram-model)), *Programmnaya Ingeneria,* 2016, vol. 7, no. 7, pp. 291—306 (in Russian).
6. **Artemov A. A.** Model' ocenki urovnja ugroz informacionnyh vyzovov planu soderzhanija informacionnogo prostranstva social'no-telekommunikacionnoj sistemy (A model for assessing the level of threats to information calls to the content plan of the information space of the socio-telecommunications system), *Informacioniy vojni*, 2015, no. 3, pp. 83—97 (in Russian).
7. **Shulenin V. P.** *Matematicheskaja statistika. CH. 1. Parametricheskaja statistika: uchebnik* (Math statistics. Part 1. Parametric statistics: a textbook), Tomsk, Publishing house of NTL, 2012, 540 p. (in Russian).
8. **Kendall M., Stewart A.** *Statisticheskie vyvody i svjazi* (Statistical conclusions and connections), Moscow, Nauka, 1973 (in Russian).
9. **Ayvazyan S. A., Enyukov I. S., Meshalkin L. D.** *Prikladnaja statistika* (Applied statistics), Moscow, Finance and Statistics, 1985, 488 p. (in Russian).