Доклад

Математическое моделирование процесса изменения содержания информационного пространства социума (Мем-грамм-модель)

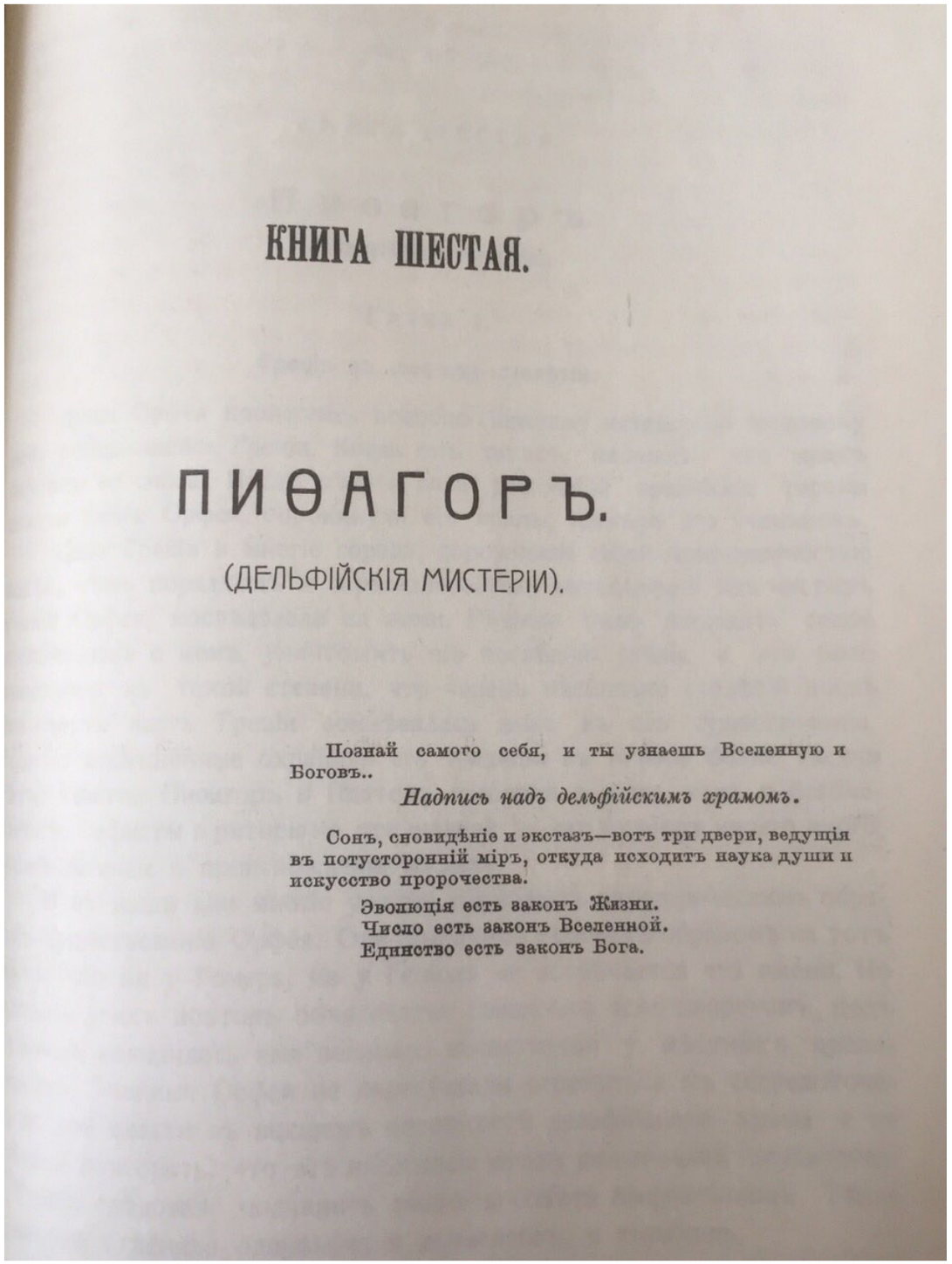
Артёмов А.А.

Семинар по социофизике, МГУ им. М.В.Ломоносова
2017 год

Вступление

<u>Эволюция</u> есть закон жизни
<u>Число</u> есть закон вселенной
<u>Единство</u> есть закон Бога

из книги «Великие посвященные»
Эдуарда Шюре



КНИГА ШЕСТАЯ.

ПИѲАГОРЪ.

(ДЕЛЬФІЙСКІЯ МИСТЕРІИ).

Познай самого себя, и ты узнаешь Вселенную и Боговъ..

Надпись надъ дельфійскимъ храмомъ.

Сонъ, сновидѣніе и экстазъ—вотъ три двери, ведущія въ потусторонній міръ, откуда исходитъ наука души и искусство пророчества.

Эволюція есть законъ Жизни.
Число есть законъ Вселенной.
Единство есть законъ Бога.

Истоки проблематики

А. Массовая информация, формирует четкие контуры моделей мира у представителей социума (Э. Ноэль–Нойман);

Б. Модели мира определяют логику поведение индивида. Совокупность индивидов со схожими моделями мира образуют социальные общности. Полярные модели мира создают конфликт между индивидами или соответствующими общностями. Управление моделями мира – является ключевым фактором управления информационным обществом (Луман, Липпман, Расторгуев).

В. Возможность математического описания изменений содержания ИП, позволит разработать научный фундамент для прогнозирования и управления ИП социума

Часть 1

Общая математическая модель

Предмет исследования

Предметом исследования, является изменение содержания информационного пространства (ИП)

Изучение данного предмета невозможно без ответов на вопросы:

A. Что такое информационное пространство социума?

B. Единицы содержания информационного пространства социума?

C. Метрика оценки изменения содержания информационного пространства социума?

Информационное пространство социума

Информация – сведения, воспринимаемые человеком или специальными устройствами как отражение фактов материального мира в процессе коммуникации (ГОСТ 7.0-99)

ИП социума – это вся совокупность информации используемая социумом

Информационное пространство социума

Социально-телекомуникационная система (субъект)

совокупность технологических средств массовой коммуникации, источников информации и социума, использующего их для обмена информацией

Информационное пространство социально-телекоммуникационной системы (объект)

часть системы знании социума, формируемая содержанием информационных сообщении, циркулирующих в СТС

Содержание информационного пространства

Актуальность исследования. Применяемые сегодня на практике и в теории оценки содержания не обладают математическим описаниям эволюции содержания всего информационного пространства (с позиции интегральной оценки).

Направление научной и практической деятельности	**Объекты**		**Единица (мера)**	
	первичные	**производные**	**качественная**	**количественная**
Массовые коммуникации	Содержание сообщения (начало XX века)	Количество просмотров или действий относительно сообщения (н. в.)	Сообщение или публикация	Охват или действие
Теория информации	Количество информации (Хартли, 1928)	Ценность информации (Харкевич, 1960)	Сигнал или событие	Бит
Когнитивная и структурная лингвистика	План выражения (Витгенштейн, Гумбольд, Фреге, 1922)	План содержания (Витгенштейн, Гумбольд, Фреге, 1922)	Слово (лексема)	Семема (смысл)
Корпусная лингвистика	n-gramm (Браун, 1988)	word2vec (Миколов, 2010)	N-грамма или набор слов	Вероятность словосочетания

Эволюция, мутация и коммуникация

Предварительные выводы международных экспертов подтверждают наш тезис о том, что MH17 сбили российские военные.Об этом на своей странице в микроблоге Twitter сообщил секретарь СНБО Андрей Парубий, подтверждает, что Ходаковский в интервью признал наличие у террористов Буков. АУДИО Предварительные выводы международных экспертов подтверждают наш тезис о том, что #MH17 сбили российские военные из российского оружия, - написал он. Напомним, что 17 июля в Донецкой области был уничтожен самолет Boeing 777 Малаqзийских авиалиний, летевший из Амстердама в Куала-Лумпур. На борту лайнера находились 298 человек, в том числе 15 членов экипажа. Во время поисковой операции были найдены 282 тела и 87 фрагментов тел.

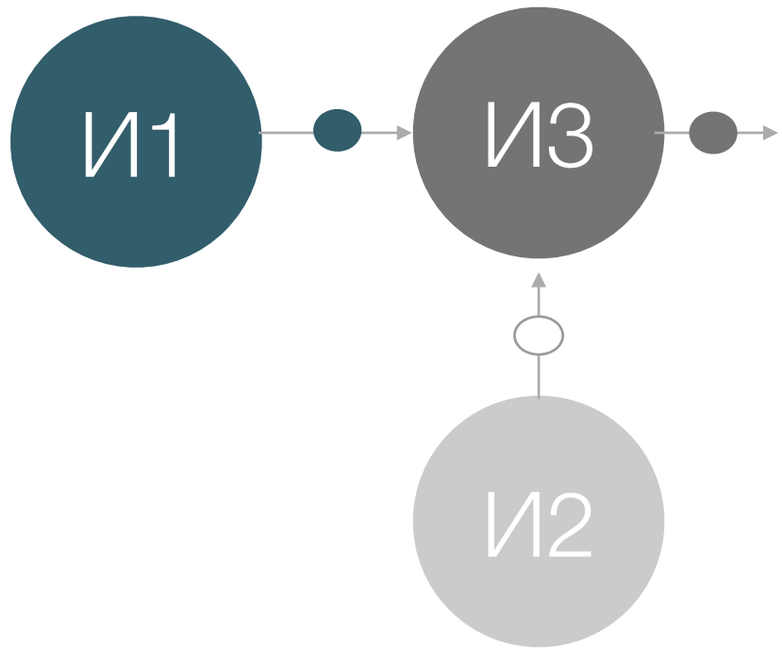
Коммуникационная операция
является естественным когнитивные фильтром
моделей знании

Редакция Ria.ru следит за обстановкой на Украине, в Донецкой и Луганской республиках, где продолжается противостояние между подконтрольными Киеву силовиками и представителями ополчения.

Содержание информационного пространства

Содержание информационного пространства определено количеством и качеством (содержанием) информационных сообщений.

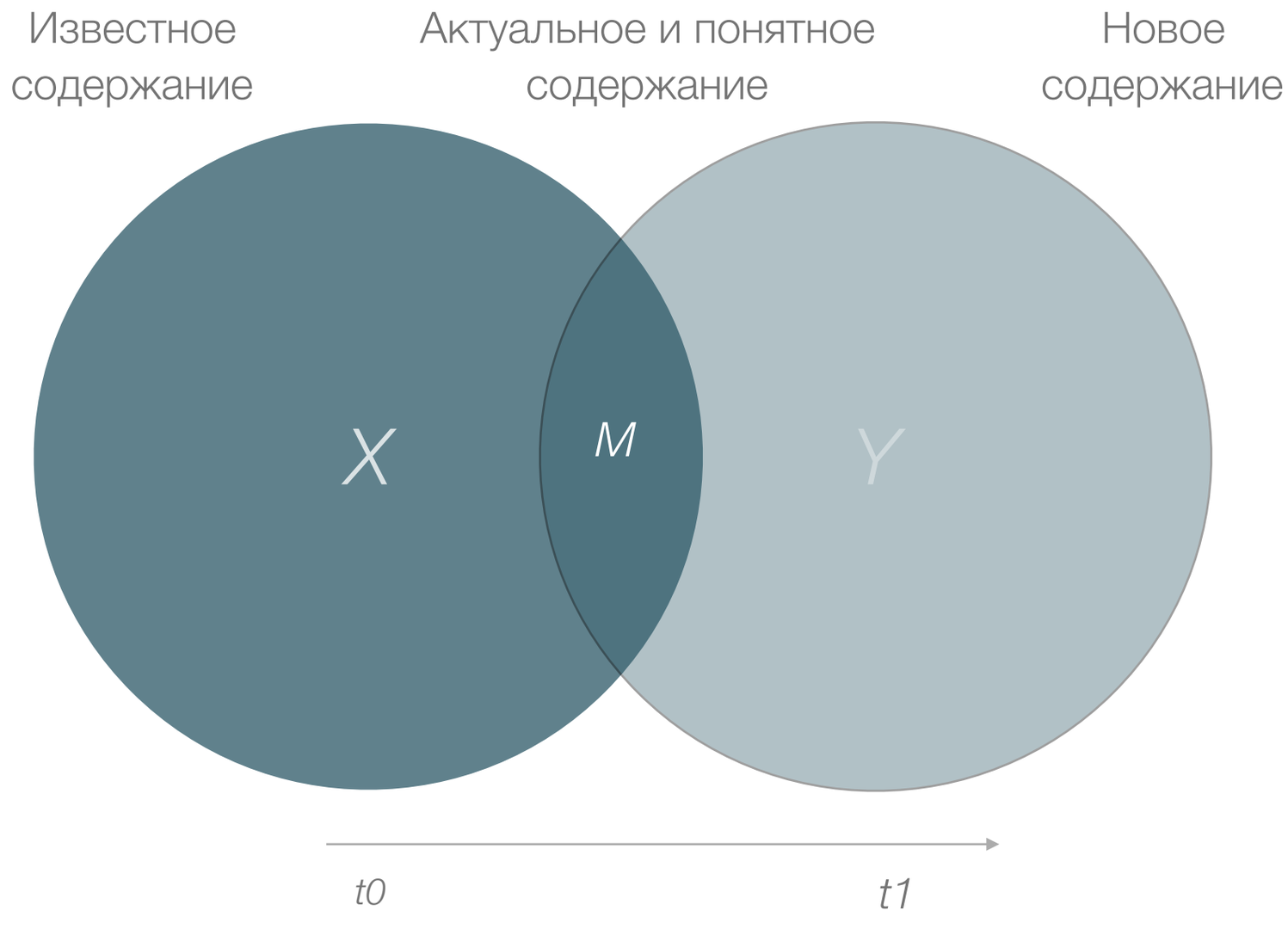
С одной стороны, ключевой особенностью речи является возможность рекурсии речевых конструкции, т.е. возможность продолжения предложений для формирования рассказа. С другой стороны, текст тем актуальнее, чем больше оно несет информации для получателя. Соответственно если текст копирует содержание другого текста - он не несет новой информации в содержании, хотя и повышает субъективную достоверность копируемого содержания. С третьей стороны, чем понятнее содержание текста, тем больше информации способен нести текст.



Таким образом описание содержания должно быть основано на интегральной характеристике, способной сочетать: актуальность, достоверность, понятность

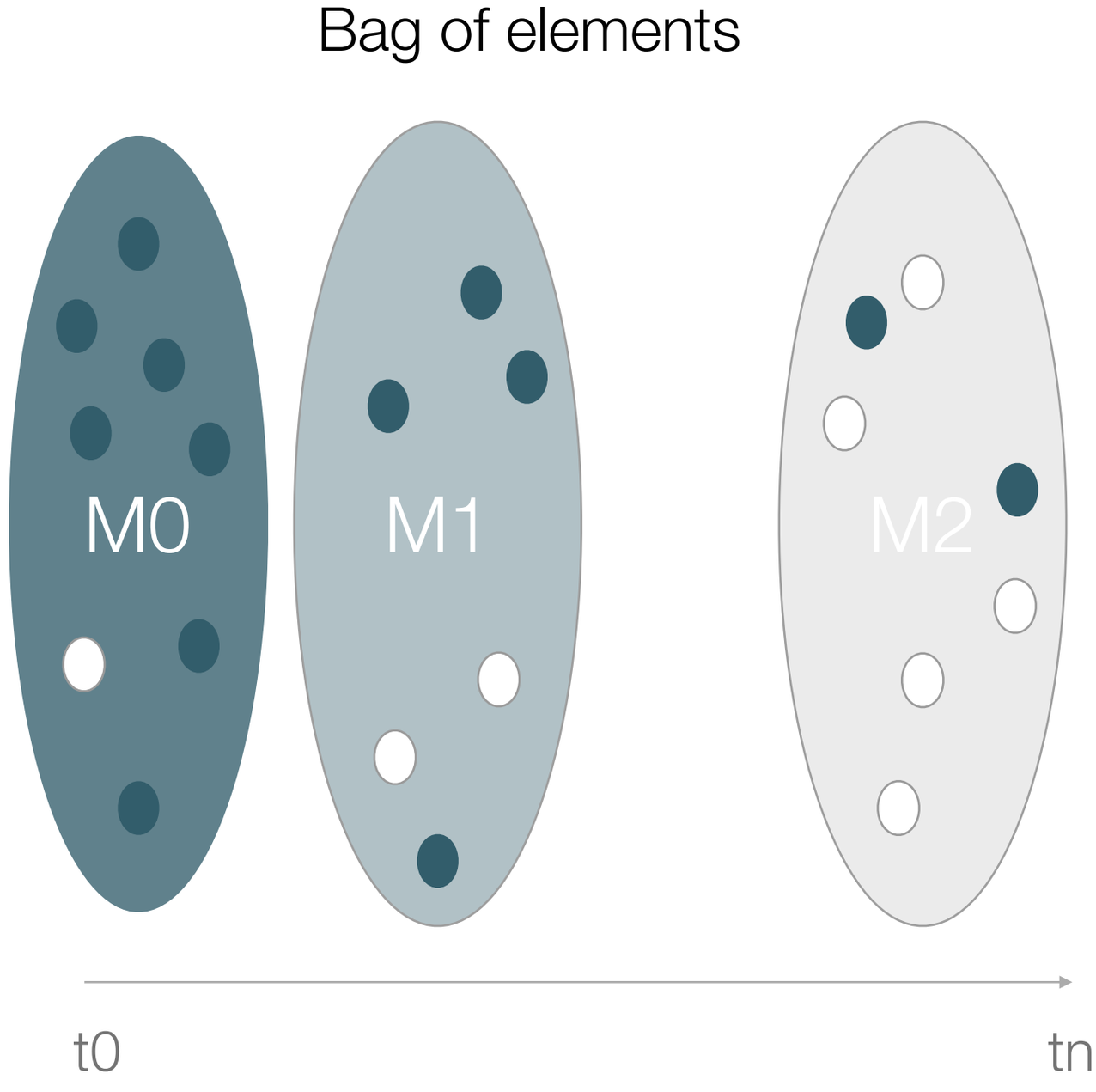
Содержание информационного пространства

Если положить, что содержание информационных сообщений может быть описано языком (Хайдеггер), то любая часть информационного пространства представляет некоторый фрагмент текста (Налимов). А совокупность текстов образуют язык СТС.



Содержание информационного пространства

Для интегрального описания содержания информационного пространства необходимо определить единицы наследственности информационного содержания



Выбор единицы наследственности информационного содержания

План содержания - организованная определённым образом область всего того, что может быть предметом языкового сообщения
Сема - минимальная, предельная единица плана содержания
Семема – единица языкового содержания (смысла), совокупность сем
Лексема – наименьшая, предельно малая единица языка, которая имеет план содержания и план выражения.

Выбор метрики наследственности информационного содержания

2 слова ≤ 1 семема

Лингвистическая практика показывает, что образование новых семем требует интервала времени T такого, что величина T/Δt - характеризующая частоту появления ИС в СТС (за интервал Δt) содержащих данную прото-семему лексемы, всегда кратно больше 1-го интервала Δt. Иными словами, для формирования нового смысла (семемы) требуется, чтобы он был воспринят коммуникаторами и неоднократно использован при коммуникации. Кроме того, возможны варианты появления и утраты нового плана содержания за незначительный период Δt, что не будет зафиксировано в семеме, но имело место быть в факте коммуникации посредством ИС СТС.

Частота семем << Частоты сообщении

Выбор метрики наследственности информационного содержания

Согласно, исследованиям группы ученых из Стэндфордского и Корнельского университетов (Jure Leskovec, 2011), мем является ключевой характеристикой в оценке изменения содержания информационных сообщении. Основываясь на данной работе, зададим признаки мема:

- Не снижение числа копий в ИП СТС;
- Продолжительный рост числа копии мема – ведет к созданию новой семы (культурной константы);
- Спад числа копии мема более заданного предела - ведет к переходу мема к набору лексем;
- Мемы распространяющиеся совместно формируют мемплексы.

Набор лексем ⇄ Мем → Новая Сема

Биграмма ≤ 1 Мем < Мемплекс

Поскольку мемы – характеризуют, с одной стороны, изменение известного плана содержания - набора лексем, а с другой, наследственность (выраженную этим набором) ИС, что в конечном счете определяет изменчивость ИП СТС – то мем, отвечает требованиям искомой метрики.

Выбор метрики наследственности информационного содержания

Термин «мем» впервые был использован в 1976 году оксфордским биологом Ричардом Доукинсом в книге «Эгоистичный ген» (Richard Dawkins, The Selfish Gene.

Генетика	**Меметика**
Ген	Мем
Клетка	Сознание
Биологический вирус	Информационный вызов (или вирус)
Набор генов	Наборы мемов (мемплексы)
Споры/микробы	Информационные сообщения
Гены и более высокие формы	Культура
Организм	ИП СТС*
Генетическая предрасположенность	Психологическая предрасположенность
Генетическая эволюция	Культурная эволюция

Мем – минимальная структурно–функциональная языковая единица наследственности плана содержания информационных сообщений.

Языковые модели

В языковых моделях предлагается функция, способная предсказывать вероятность того, что определенная последовательность слов является допустимой для данного языка.

Модель на N-граммах

$$\mathrm{P}(\mathrm{w}_i = w_{n-1}\, w_n) = \frac{N\,(\mathrm{w}_i = w_{n-1}\, w_n)}{\sum_i N\,(\mathrm{w}_i)}$$

Модель на векторах word2vec

$$\mathrm{P}(\mathrm{w}_i = \bar{w}_i\, w_c) = \frac{e^{u_{w_c}^T * V_{w\,i}}}{\sum_C e^{u_{w_c}^T * V_{w_i}}}$$

Эти и другие подобные им модели 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.

Для оценки изменения ИП, необходим подход, позволяющий учитывать оценку изменения языковой модели представления знаний (сформированной на основе словосочетаний w корпуса текстов T) для прогноза вероятности использования словосочетания w в текстах будущего периода.

Математическая модель

Решать задачу 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 языковой модели в целом и последующим учетом этого изменения для каждого словосочетания в будущем. Основываясь на опыте [15,20], в котором рассматривается модель для отслеживания изменения мемов – как устойчивых «мутирующих» словосочетаний (meme tracker, стэндфордская модель) и работе [21], где описана компьютерная модель процесса эволюции культуры (memes and variations, MAV), автор пришел к выводу о целесообразности выбора эволюционного подхода «от частного к общему» при решении задачи оценки изменения и прогноза содержания информационного пространства. Исходя из данного выбора, были сформулированы 2 подзадачи для решения целевой задачи оценки изменения языковой модели:

1. Определить количественную меру отличия одной языковой модели вида $\langle C_{T1}, P(w_{i1})\rangle$ относительно объектов (словосочетания) другой модели вида $\langle C_{T2}, P(w_{i2})\rangle$;
2. Разработать алгоритм для прогноза допустимости (вероятности) $P(w_i; dt)$ словосочетания w для языка в будущий момент времени (языковой модели вида $\langle C_T, P(w_i)\rangle$ за интервал времени dt).

Решением поставленных задач явилось создание мем-грамм-модели.

Формализция ИП

1. Набор текстов

$$T_i^{\tau}\ \{w_1, w_2, \dots w_k\}. \qquad \{T_1^{\tau}, T_2^{\tau}, \dots T_i^{\tau}\}$$

2. Коолекция корпусов

$$C = \begin{pmatrix} C_1\ \left\{w_1^1, w_2^1, \dots w_{k1}^1 \mid \forall T_i^{t_n = t_p}\right\} \\ C_2\ \left\{w_1^2, w_2^2, \dots w_{k2}^2 \mid \forall T_i^{t_n = t_p}\right\} \\ \dots \\ C_p\ \left\{w_1^p, w_2^p, \dots w_k^p \mid \forall T_i^{t_n = t_p}\right\} \end{pmatrix}$$

$$C^P = \begin{pmatrix} \langle C_1, P_1(w_{k1})\rangle \\ \langle C_1, P_1(w_{k2})\rangle \\ \dots \\ \langle C_p, P_p(w_k)\rangle \end{pmatrix}$$

3. Коллекция мемов

$$C_p\colon \varphi\big(\langle m_1 = w_1, m_2 = w_2, \dots m_j\rangle \mid \forall m_j = w_k\big) \Rightarrow$$
$$M_p = Supp(C_{p-1} \cap C_p),\ j \in \mathbb{N}, 0 < j \leq k,\ M_p$$

$$M^{\varphi} = \begin{pmatrix} M_1 = \emptyset \\ C_1 \cap C_2 \overset{\varphi}{\Rightarrow} M_2\ \{m_1^1, m_2^1, \dots m_{j1}^1\} \\ C_2 \cap C_3 \overset{\varphi}{\Rightarrow} M_3\ \{m_1^2, m_2^2, \dots m_{j2}^2\} \\ \dots \\ C_{p-1} \cap C_p \overset{\varphi}{\Rightarrow} M_p\ \{m_1^p, m_2^p, \dots m_j^p\} \end{pmatrix}$$

Математическая модель

Определение 1. Информационным пространством (ИП), на котором можно задать метрику, будем называть пару коллекций $\langle C^P, \mathrm{M}^{\varphi} \rangle$, где C^P - коллекция языковых (н-грамм) моделей $\langle C_p, P(w_k) \rangle$ в дискретные моменты времени p, M^{φ} – коллекция множеств мемов, общих элементов C_{p-k} и C_p. $P(w_k)$ - вероятность словосочетания или вектора из слов w_k.

Отметим, что в рамках данной работы, рассматриваются языковые н-грамм модели (для случая биграмм), где вероятность мема, представленного биграммой, определяется как отношение частоты встреч биграммы мема, к числу всех биграмм.

$$\mathrm{P}\left(m_i^p\right) = \frac{\sum_{\forall\, m_i^p = \mathrm{w}_k^p \in \mathrm{C_p}} 1}{\sum_w N\left(\mathrm{w}_j^p\right)} = \frac{N\left(\mathrm{w}_j^p = m_i^p\right)}{|C_p|} \qquad 0 \le \mathrm{P}\left(\mathrm{w}_j^p \ne m_i^p\right) \le \left(1 - \sum\nolimits_p \mathrm{P}\left(\mathrm{w}_j^p = m_i^p\right)\right)$$

$$\sum\nolimits_j N\left(\mathrm{w}_j^p = m_i^p\right) + \sum\nolimits_j N\left(\mathrm{w}_j^p \ne m_i^p\right) = |C_p|$$

Математическая модель

С применением моделей на н-граммах решается задача предсказания вероятности того, что определенная последовательность слов является допустимой для данного языка $\mathrm{P}(\mathbf{слово}_\mathrm{n}|\mathbf{слово}_{\mathrm{n}-1})$. В мем-грамм-модели решается иная задача – предсказать, что допустимая для языка $\mathrm{C_p}$ последовательность слов $w_{\mathrm{n}-1}\, w_\mathrm{n}$ в текущий момент времени p, будет допустимой для языка $\mathrm{C_{p+1}}$ в следующий момент времени p+1 или, что равнозначно, будет являться мемом $\mathrm{m_q^{p+1}} = w_{\mathrm{n}-1}\, w_\mathrm{n}$ для языков $\mathrm{C_p}$ и $\mathrm{C_{p+1}}$ в следующий момент времени.

$$\mathrm{P}\left(\mathrm{w}_j^{p+1} = m_i^p\right) = \frac{N\left(\mathrm{w}_j^{p+1} = m_i^p\right)}{\left|C_{p+1}\right|}$$

1-?

2-?

Математическая модель

1. Количество элементов языковых моделей постоянно и примерно равное, то есть $|C_1| \approx |C_2| \approx \cdots |C_p|$, такое условие можно довольно просто обеспечить – контролем мощности мультимножеств C_p при формировании его из н-грамм текстов.
2. Разнообразие мемов $|M_k|$ пропорционально «размеру» языковой модели ИП $|C_k|$, то есть $\frac{|C_q|}{|M_q|} \approx const$. Это условие выглядит логично и согласуется [24] с Законом необходимого разнообразия Эшби: «разнообразие (энтропию) управляемого можно понизить не более чем на величину количества информации в управляющей системе об управляемом, которое равно разнообразию (энтропии) управления за вычетом потери информации от неоднозначного управления». Иначе говоря, значительное разнообразие воздействующих на большую и сложную систему возмущений требует адекватного им разнообразия её возможных состояний

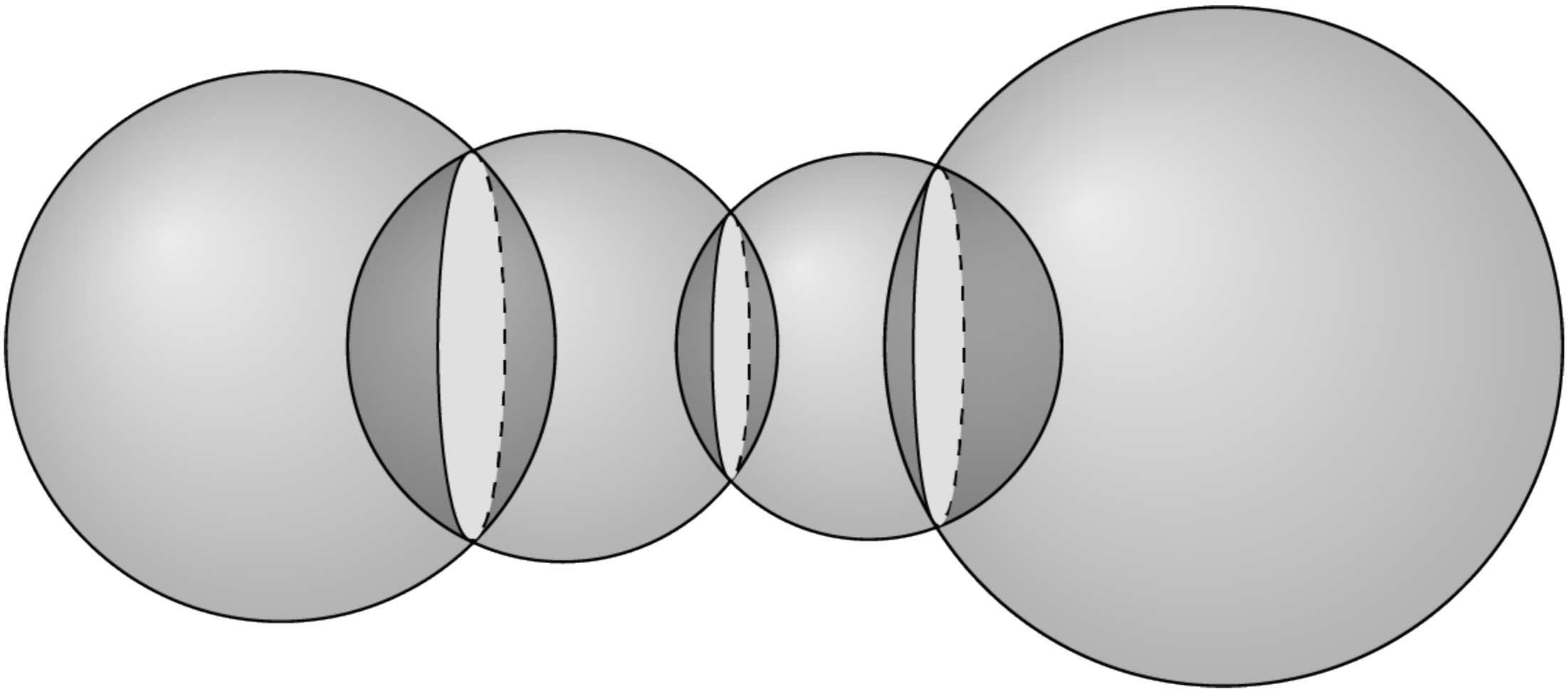
Математическая модель

Определение 2. Если задана функция f такая, что ставит в соответствие каждому моменту времени $p \in \mathbb{N}$, значение $\xi \in \mathbb{R}_0$ так, что $\left|\frac{|C_p|}{|M_p|} - \frac{|C_{p-1}|}{|M_{p-1}|}\right| \leq \xi$ будем говорить, что определен закон изменения разнообразия f в ИП.

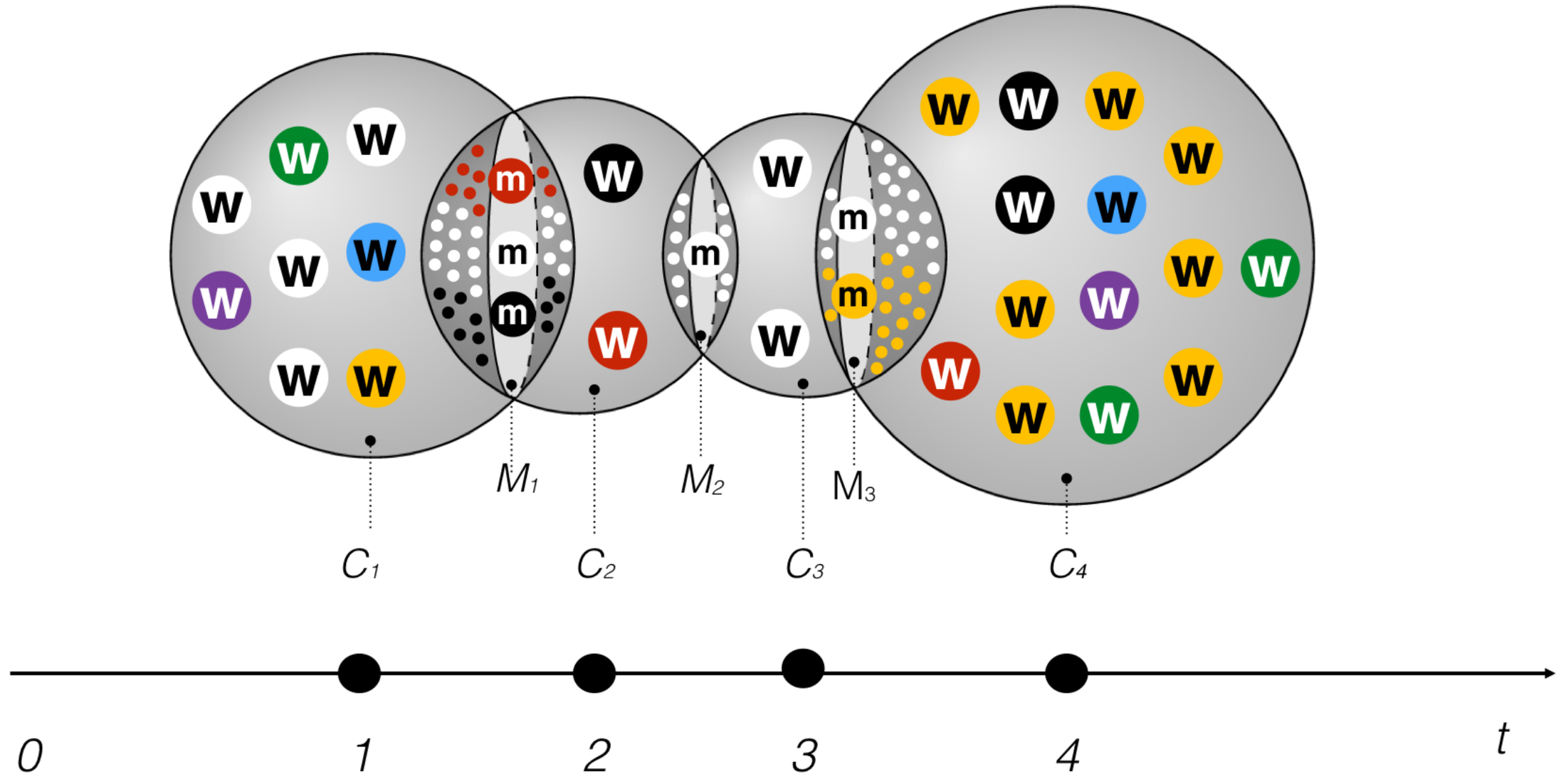
Определение 3. Мультимножества C' и C будем называть подобными (или частично S-эквивалентными) $C' \underset{\sim}{}^{k^M} C$ относительно всех элементов множества $M \neq \emptyset$ с коэффициентом подобия k, если выполняется условие: $\exists k$: $\forall \mathrm{w}_i \in C, \mathrm{w}_j \in C', m_r \in M$, так что $\mathrm{N}(\mathrm{w}_i = m_r) = k * \mathrm{N}(\mathrm{w}_j = m_r)$, где N - функция кратности, $k \in \mathbb{R}$.

Аксиома 1. В информационном пространстве сохраняется закон изменения разнообразия f, так что $\alpha > 2\xi$ и $\alpha - 2\xi \leq \frac{|C_q|}{|M_q|} \leq \alpha + 2\xi$, $\forall q \in \{1,2,..\mathrm{p}\}$, где, $\{1,2..\mathrm{p}\}$ – множество индексов p-моментов времени, используемых в коллекции C^P, $\xi = \lim_{p\to\infty} \frac{\sum_p \left|\frac{|C_p|}{|M_p|} - \frac{|C_{p-1}|}{|M_{p-1}|}\right|}{P}$, $\alpha = \lim_{p\to\infty} \frac{\sum_p \frac{|C_p|}{|M_p|}}{P}$

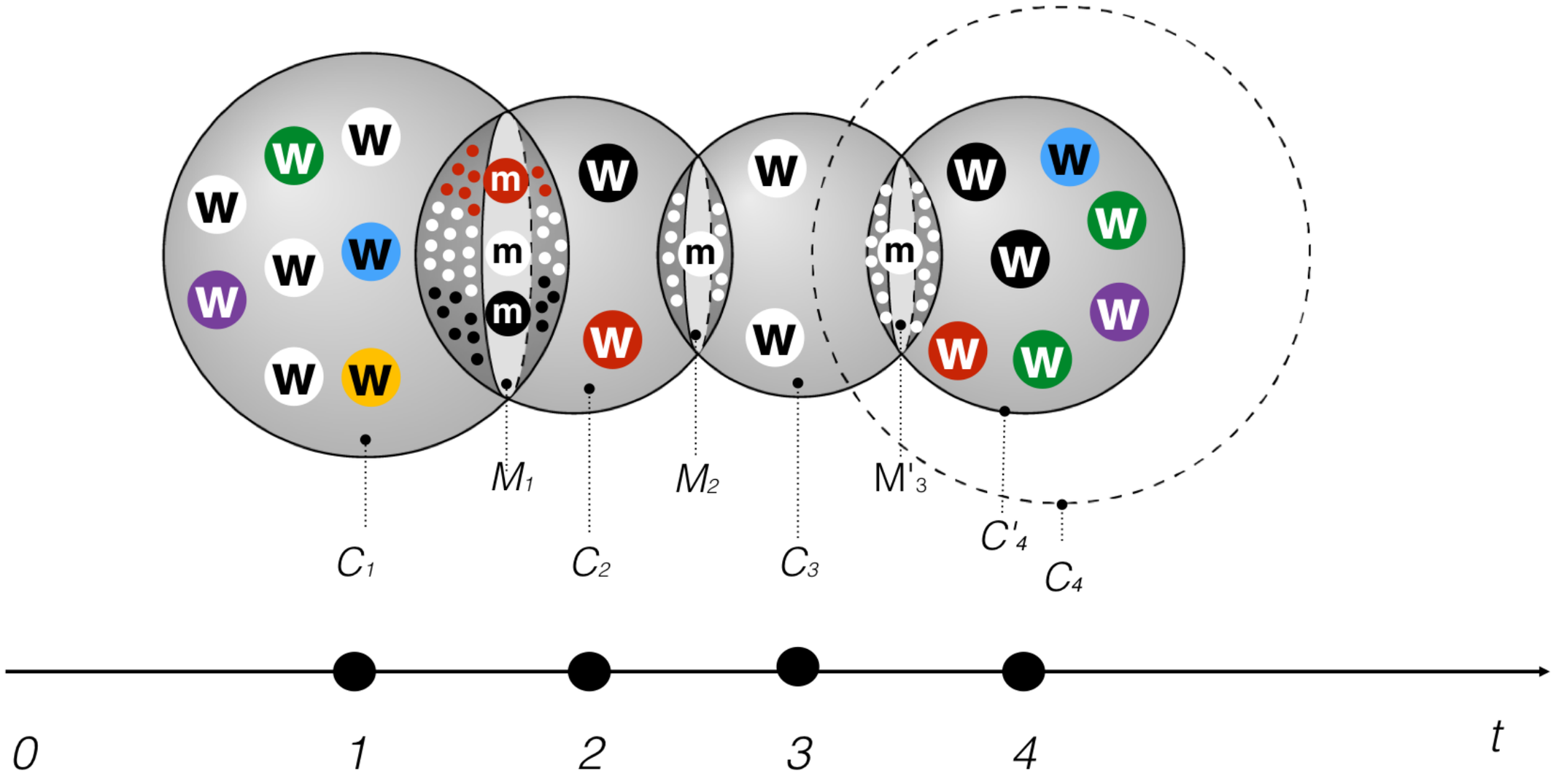
Топологическая задача



Динамика изменения ИП



Математическая модель



Эволюция

$$2 \cdot (\max htgC_q - 1) < \alpha \leq \max / C_p /$$

Революция

$$\alpha > \max / C_p /$$

$/ C_p / = \dim C_p$ - размерность мультимножества C_p

Математическая модель

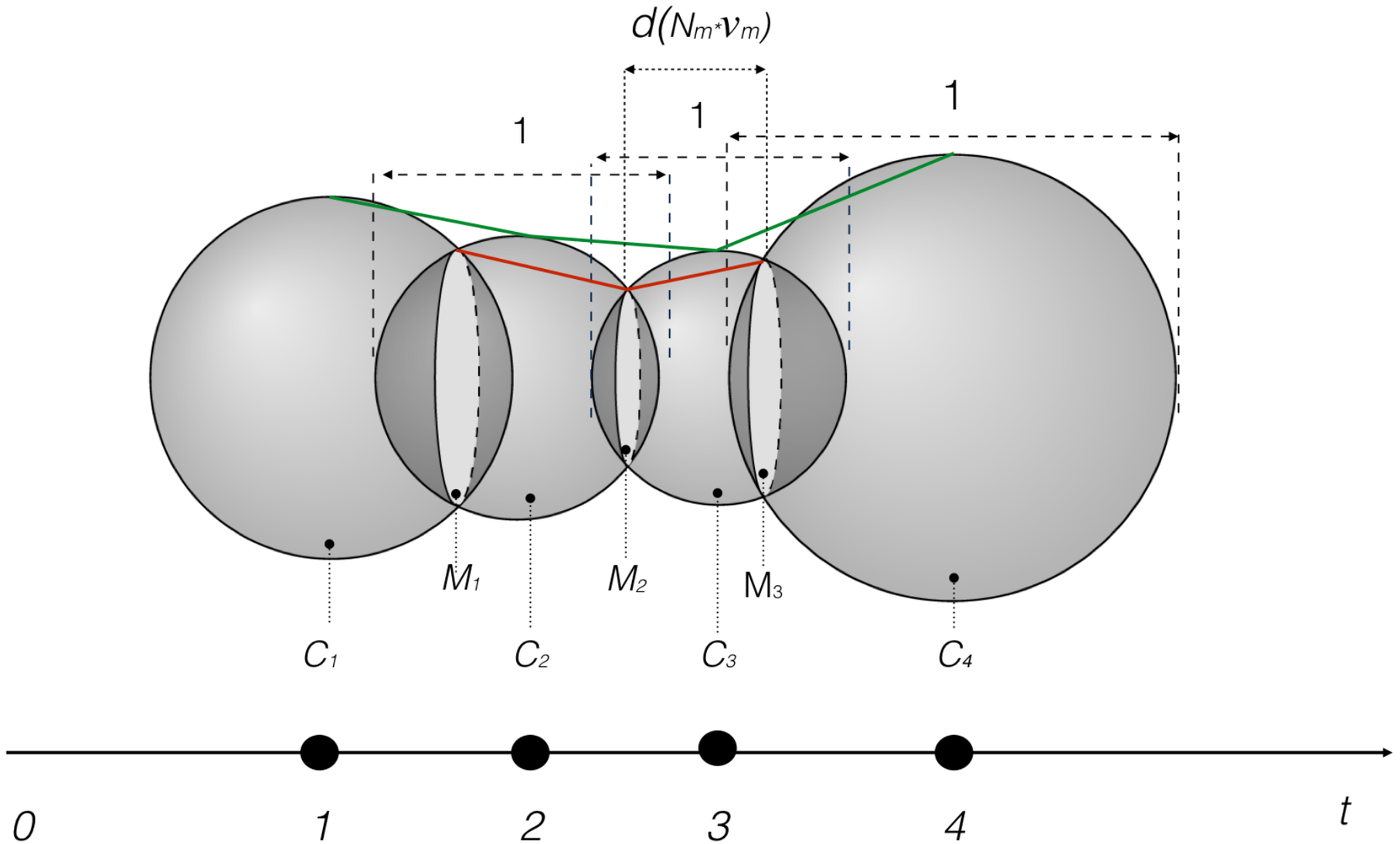
Утверждение 3. Если $C' \underset{\sim}{}^{k^C} C$, $C \neq \emptyset$, $M = Supp(C'' \cap C) \neq \emptyset$, $M' \subseteq C$, $\frac{|C|}{|M|} \approx \alpha$, то для выполнения $\frac{|C'|}{|M'|} \approx \alpha$, необходимо чтобы $k^C = \frac{|M|}{|M'|}$.

Доказательство. Предположим $k^C \neq \frac{|M|}{|M'|}$, тогда следует $\frac{|C'|}{|M'|} \neq \frac{|C|}{|M|} \approx \alpha$, $k^C \neq \frac{|C|}{|C'|}$. Подобие $C' \underset{\sim}{}^{k^C} C$, означает $N(\mathrm{w}_i) = k^{C_{p-1}} * N(\mathrm{w}_j)\, \forall \mathrm{w}_i \in C, \mathrm{w}_j \in C'$ $|C| = \sum_j N(\mathrm{w}_j)$, $|C'| = \sum_i N(\mathrm{w}_i) \Rightarrow |C| = k^C * |C'_p|$ то есть $k^C = \frac{|C|}{|C'|}$, что противоречит предположению, следовательно, $k^C = \frac{|C|}{|C'|} = \frac{|M|}{|M'|}$, а так как $\frac{|C|}{|M|} \approx \alpha$, то $\frac{|C'|}{|M'|} = \frac{|C|}{|M|} \approx \alpha$.■

Теорема1. В ИП, где сохраняется закон разнообразия, вероятностная мера общих элементов текущего и следующего ИП не зависит от общего числа словосочетании в следующем ИП.

$$\mathrm{P}\left(\mathrm{w}_j^{p-1} = m_i^p\right) \approx \frac{N(\mathrm{w}_j^{p-1} = m_i^p)}{\alpha \cdot |M'_p|}$$

Математическая модель



$$\vec{\mathcal{V}}\left(m_i^{z,\Delta z}\right) = \frac{\mathrm{P}(m_i^z) - \mathrm{P}\left(\mathrm{w}_j^{z-1} = m_i^z\right)}{\Delta z = 1}$$

$$F_{z,\Delta z}^{m_i} = \sum \vec{f}_{z,\Delta z}^{m_i} = \frac{\Delta(\mathrm{N}(m_i^z)\vec{\mathcal{V}}\left(m_i^{z,\Delta z}\right))}{\Delta z}$$

Математическая модель

Определение 4. Информационным пространством (ИП) с метрикой будем называть тройку $\left(\langle \mathrm{C}^{\mathrm{P}}, \mathrm{M}^{\varphi}\rangle, \alpha^{\mathrm{IS}}\right)$, где C^{P} - коллекция языковых моделей $\langle \mathrm{C_p}, \mathrm{P(w_k)}\rangle$ в дискретные (Марковские) моменты времени p, M^{φ} – коллекция множеств M общих элементов $\mathrm{C_{p-k+1}}$ и $\mathrm{C_{p-k}}$, α^{IS} – отображение, ставящее в соответствие двум множествам M коллекции M^{φ} некоторое вещественное число, определяющее меру изменения текущей языковой модели $\mathrm{C_{p-k+1}}$ (относительно предыдущей $\mathrm{C_{p-k}}$ и следующей языковой модели $\mathrm{C_{p-k+2}}$).

$$\alpha^{IS} = \left|F_{p,\Delta p}^{\text{ИП}+} - F_{p,\Delta p}^{\text{ИП}-}\right|$$

Где сила создания и уничтожения элементов наследственности содержания ИП

$$F_p^{\text{ИП}+} = \sum\nolimits_i \left(F_{p,\Delta p}^{m_i} | \forall F_{p,\Delta p}^{m_i} \geq 0 \right) \qquad F_p^{\text{ИП}-} = \sum\nolimits_i \left(F_{p,\Delta p}^{m_i} | \forall F_{p,\Delta p}^{m_i} < 0 \right)$$

Математическая модель

Неизвестным остается $N\left(\mathrm{w}_{\mathrm{i}}^{\mathrm{p+1}} = m_i^p\right)$. Таким образом, необходимым и достаточным условием существования функции вероятности $\mathrm{P}\left(m_i^{P+1}|m_j^P\right)$ является определение функции или алгоритма Λ, который позволит прогнозировать количество $N\left(\mathrm{w}_{\mathrm{i}}^{\mathrm{p+1}} = m_i^p\right)$ копии мема $\mathrm{m}_{\mathrm{q}}^{\mathrm{p+1}} = m_j^P = \mathcal{w}_{n-1}\,\mathcal{w}_n = \mathrm{w}_{\mathrm{n}} \in \mathrm{C}_{\mathrm{p}}$ в будущий период с заданной достоверностью θ.

$$\mathrm{P}\left(m_i^{P+1}|m_j^P = \mathcal{w}_{n-1}\,\mathcal{w}_n\right) = \frac{\mathrm{P}\left(\mathcal{w}_{n-1}\,\mathcal{w}_n|m_j^{P+1}\right)\mathrm{P}\left(m_j^{P+1}\right)}{\sum_i \mathrm{P}\left(\mathcal{w}_{n-1}\,\mathcal{w}_n|m_i^{P+1}\right)\mathrm{P}\left(m_i^{P+1}\right)} = \frac{\mathrm{P}\left(m_j^{P+1}\right)}{\sum_i \mathrm{P}\left(m_i^{P+1}\right)} \approx \frac{\frac{N\left(\mathrm{w}_{\mathrm{j}}^{\mathrm{p+1}} = m_j^p\right)}{\alpha \cdot |M'_{p+1}|}}{\sum_j \frac{N\left(\mathrm{w}_{\mathrm{j}}^{\mathrm{p+1}} = m_j^p\right)}{\alpha \cdot |M'_{p+1}|}} \approx \frac{N\left(\mathrm{w}_{\mathrm{j}}^{\mathrm{p+1}} = m_i^p\right)}{\sum_j N\left(\mathrm{w}_{\mathrm{j}}^{\mathrm{p+1}} = m_i^p\right)}$$

Определение 5. Мем-грамм-модель задана, если в информационном пространстве $(\langle C^P, \mathrm{M}^{\varphi}\rangle, \alpha^{IS})$ определена $\Lambda^{\theta,\varepsilon}$ – функция (или алгоритм) прогноза числа экземпляров $N\left(\mathrm{w}_{\mathrm{i}}^{\mathrm{p+1}} = m_i^p\right)$ мема $\mathrm{m}_{\mathrm{i}}^{\mathrm{p+1}} = \mathrm{m}_{\mathrm{j}}^{\mathrm{p}} = \mathrm{w}_{\mathrm{q}}^{\mathrm{p}} \in \mathrm{C}_{\mathrm{p}} \subset C^P$, $\mathrm{m}_{\mathrm{j}}^{\mathrm{p}} \in \mathrm{M}_{\mathrm{p}} \subseteq \mathrm{M}^{\varphi}$, θ – оценка достоверности прогноза, ε - максимальная ошибка прогноза. В соответствии с данными дополнениями необходимым условием для задания мем-грамм-модели является задание четверки элементов $\left(\langle C^P, \mathrm{M}^{\varphi}\rangle, \alpha^{IS}, \Lambda^{\theta,\xi}\right)$.

Математическая модель

Автор не смог найти приема для разработки чисто теоретического математического метода определения функции $\Lambda^{\theta,\varepsilon}$, да и вряд ли это возможно в отсутствии гипотез, основанных на изучении процессов изменения реального ИП. Поэтому сначала был разработан инструментарий (программное обеспечение) для сбора и обработки реальных данных – 3000 публикаций русскоязычного сегмента Интернета за июнь-август 2014 года [8], а затем на их основе были созданы н-грамм-модели, коллекция которых сформирована как объект изучения – ИП русскоязычного сегмента Интернета (ИПР) [23]. Изучение процессов, происходящих в ИПР, позволили сформировать метрику силы информационного воздействия и подтвердить пригодность концепции меметики для описания изменений ИП. Но главное, оно позволило сформировать гипотезу «Об эволюции популяции мемов», которая в последствии подтвердилась в виде доказанной теоремы «О плотности вероятности функции числа копии мемов» и алгоритма $\Lambda^{\theta,\xi}$. Основой для подтверждения гипотезы, доказательства теоремы и определения алгоритма является разработанная статистическая модель прогнозирования числа копий мема в следующий период, которая представлена в следующей статье автора.

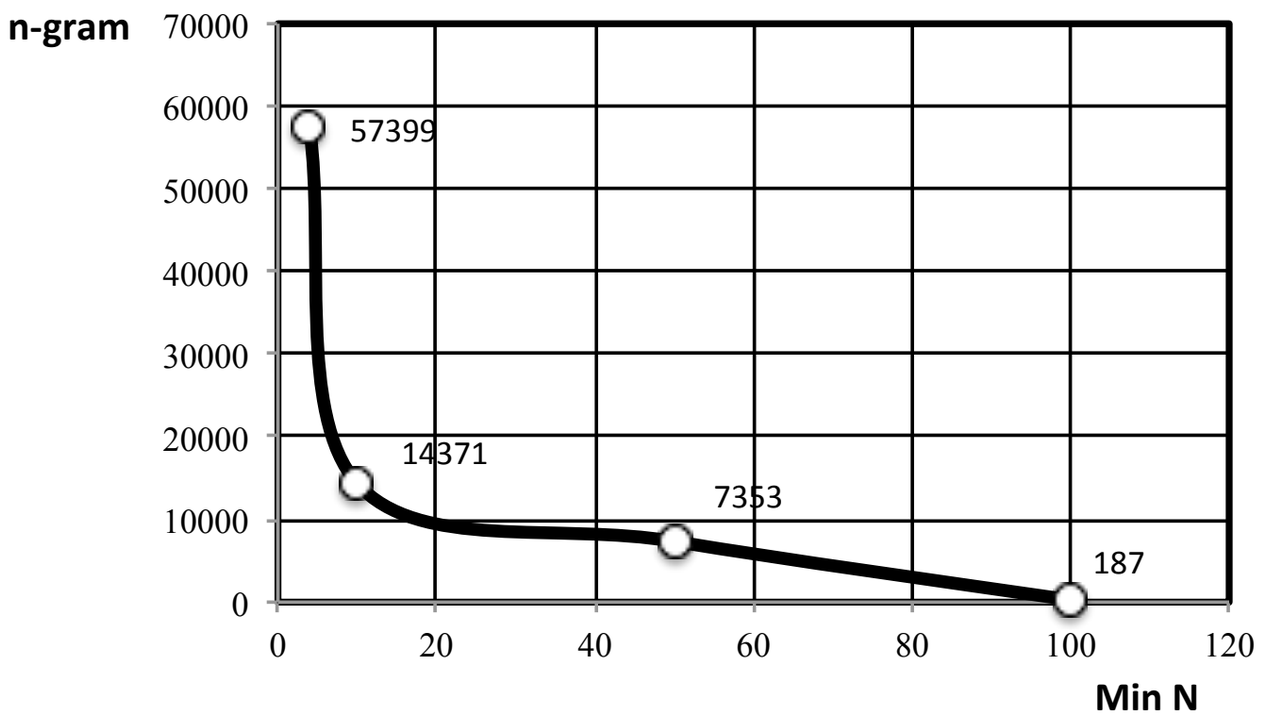
Меметический алгоритм (би-эволюционный)

Алгоритм2. МА мем-грамм-модели

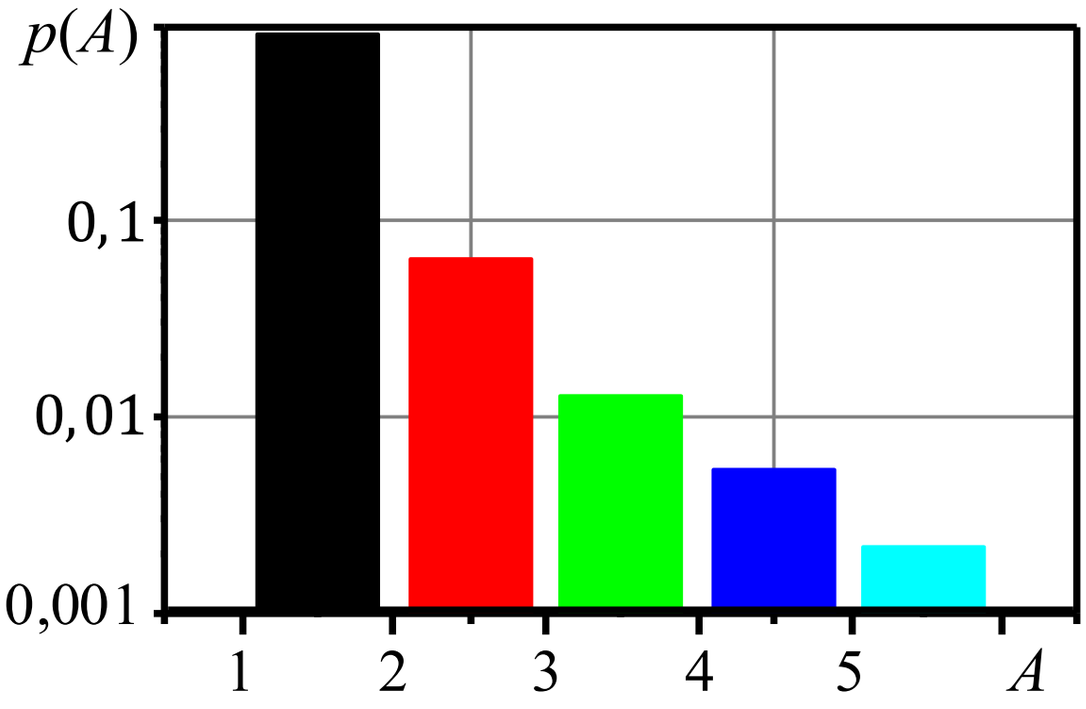
1. ***Вход***: $T_{t_0-n*\Delta t}$ // *тексты публикации за n предыдущих дискретных интервалов времени;*
2. ***Выход***: $\vec{M}_{t_0 \mp z*\Delta t}$// *наследственность содержания ИП через z дискретных интервалов времени (знак зависит от рассмотрения в перспективе или ретроспективе относительно выбранного t, z>n);*
3. *T ⇒ W // Формирование мультимножества n-грамм W, из слов текстов T*
4. *T ⇒ M // определение мемов из мультимножества n-грамм*
5. *<N_m,G_m,A_m> // оценка характеристик «приспособленности» мема*
6. *F=F_m (G,A(t)) // оценка силы информационного воздействия на мем;*
7. *m* = лучший m // выбор мемов с лучшими характеристиками <N,G,A> через* $t \mp z * \Delta t$
8. *F* = F (m*) // параметры «приспособленности» у лучших элементов наследственности*
9. ***While*** *StopCondition не выполнено do*
10. *(m_1 ... $m_к$) = ParentsSelection (W) //* $k \geq 2$, *выбор "родителей";*
11. *m'= Recombination (m_1 ... mк) // получение представителей «потомства» ;*
12. *m = OffspringImprovement (m') // улучшение потомства, обычно путем локального поиска;*
13. *W = PopulationUpdate (m, W) //популяция обновляется в соответствии с правилом разнообразия (к примеру, пропорции наследственности и изменчивости);*
14. $\alpha_t^{ИП}$ *(m*, f*) = BestSolutionUpdate (m*, f*, w) //выбор наилучшего решения и запись его значения;*
15. ***End while***

Результаты экспериментов

Всего было собрано 3 000 статей за 3 месяца, ~1,2 млн слов, по ним было выявлено 1 038 779 уникальных n-грамм. Для обеспечения условия статистической значимости были отобраны 37 716 n-грамм, у которых количество копий за 92 дней не менее 5.



Распределение н-грамм по числу копии



Распределение мемов по агрессивности

Результаты экспериментов

Мем	A	F	N1	N2	W1	W2	Дата
санкция против	1,08200	0,00071	39,00000	112,00000	0,00007	0,00014	20.06.14
санкция против	1,34600	0,00069	34,00000	113,00000	0,00006	0,00014	19.06.14
санкция против	2,02300	0,00069	27,00000	119,00000	0,00005	0,00015	18.06.14
санкция против	2,71800	0,00060	21,00000	124,00000	0,00004	0,00016	17.06.14
санкция против	2,69300	0,00054	19,00000	123,00000	0,00004	0,00016	16.06.14
гуманитарный помощь	1,02800	0,00049	37,00000	97,00000	0,00005	0,00010	27.06.14
санкция против	2,11900	0,00048	19,00000	118,00000	0,00005	0,00015	15.06.14
гуманитарный помощь	1,02100	0,00047	36,00000	95,00000	0,00005	0,00010	26.06.14
санкция против	1,85700	0,00044	19,00000	113,00000	0,00005	0,00014	14.06.14
санкция против	1,85000	0,00039	17,00000	111,00000	0,00005	0,00014	13.06.14
санкция против	1,79300	0,00033	15,00000	108,00000	0,00005	0,00014	12.06.14
барак обама	1,26100	0,00031	22,00000	79,00000	0,00004	0,00010	17.06.14
санкция против	1,38500	0,00028	15,00000	104,00000	0,00005	0,00013	11.06.14
барак обама	1,11900	0,00027	21,00000	78,00000	0,00005	0,00010	16.06.14
президент сша	1,00900	0,00022	21,00000	67,00000	0,00004	0,00008	17.06.14
вводить санкция	2,57500	0,00021	14,00000	71,00000	0,00002	0,00008	21.06.14
член экипаж	1,18800	0,00021	20,00000	62,00000	0,00004	0,00008	19.06.14
вводить санкция	1,53100	0,00021	19,00000	66,00000	0,00003	0,00007	24.06.14
член экипаж	1,60800	0,00021	14,00000	76,00000	0,00004	0,00010	14.06.14
вводить санкция	3,34900	0,00021	12,00000	72,00000	0,00002	0,00009	20.06.14

Расчет характеристик для дальнейшего для анализа, пример с Δt=30 дней

Результаты экспериментов

Мем	F, ЕИВ
донецкий область	0,00039
~~риа новость~~	~~0,00037~~
~~июн риа~~	~~0,00033~~
~~июн риа новость~~	~~0,00033~~
член экипаж	0,00026
зона ато	0,00026
народный республика	0,00026
сбивать самолет	0,00025
самый дело	0,00024
украинский военный	0,00021
местный житель	0,00021
крушение самолет	0,00021
владимир путин	0,00020
сей пора	0,00019
министр внутренний	0,00015

ТОП 15 мемов, Δt=1 день

N-gramma	Число копии
член экипаж	**23**
амстердам куалалумпур	**19**
донецкий область	**19**
крушение малайзийский	19
пассажирский самолет	18
украинский армия	18
российский войско	17
установка бук	17
пассажир член	16
самолет сбивать	**16**
малайзийский авиалиния	16
сбивать самолет	16
крушение самолет	15
небо над	15
посольство киев	15
комплекс бук	14

ТОП 15 N-грамм, Δt=1 дней

Зачеркнутый текст - мемы/N-gram, которые попали в список в результате ошибки разбора содержания XTML кода.

Выделенный текст *- N-gram, которые встречаются в разные интервалы*

Результаты экспериментов

Мем	F, ЕИВ
вооруженный сила	0,00073
донецкий область	0,00070
владимир путин	0,00065
петр порошенко	0,00063
сей пора	0,00058
крушение малайзийский	0,00057
член экипаж	0,00053
тот кто	0,00050
народный республика	0,00049
~~риа новость~~	~~0,00049~~
восток украина	0,00048
боевой действие	0,00045
территория украина	0,00044
президент рф	0,00044
президент украина	0,00043
луганский область	0,00042

ТОП 15 мемов, Δt=4 дня

N-gramma	Число копии
тот кто	**548**
вооруженный сила	**535**
местный житель	**474**
~~июла риа~~	~~459~~
зона ато	458
крушение малайзийский	**452**
украинский силовик	426
~~июла риа новость~~	~~412~~
донецкий область	**399**
сила украина	378
амстердам куалалумпур	**375**
член экипаж	**359**
пассажирский самолет	**339**
малайзийский авиалиния	325
антитеррористический операция	**318**
самолет сбивать	**266**

ТОП 15 N-грамм, Δt=4 дня

Результаты экспериментов

Мем	F, ЕИВ
санкция против	0,00073
гуманитарный помощь	0,00070
барак обама	0,00065
президент сша	0,00063
член экипаж	0,00058
гуманитарный конвой	0,00057
вводить санкция	0,00053
крушение малайзийский	0,00050
зона ато	0,00049
советский союз	0,00049
добровольческий батальон	0,00048
самолет сбивать	0,00045
~~июла риа~~	~~0,00044~~
санкционный список	0,00044
гуманитарный груз	0,00043
~~июла риа новость~~	~~0,00042~~

ТОП 15 мемов, Δt=30 дней

N-gramma	Число копии
вооруженный сила	**3880**
тот кто	**3229**
местный житель	**2439**
украинский силовик	2260
июла риа	2092
вооруженный сила украина	2048
сила украина	1887
сей пора	1763
зона ато	1753
~~июла риа новость~~	~~1678~~
антитеррористический операция	**1667**
донецкий область	1426
украинский армия	1354
про тот	1347
крушение малайзийский	**1313**
область украина	1211

ТОП 15 N-грамм, Δt=30 дней

Результаты экспериментов

H-граммы отобранные по принципу частотности приводят к повторяемости описания содержания, что в итоге дает не богатство описания картины, а бедность.

Список по ТОП15 H-грамм описывающих ИП за июнь-июль-август 2015

амстердам куалалумпур, антитеррористический операция, вооруженный сила, вооруженный сила украина, донецкий область, зона ато, июла риа, июла риа новость, комплекс бук, крушение малайзийский, крушение самолет, малайзийский авиалиния, местный житель, небо над, область украина, пассажир член, пассажирский самолет, посольство киев, про тот, российский войско, самолет сбивать, сбивать самолет, сей пора, сила украина, тот кто, украинский армия, украинский силовик, установка бук, член экипаж **(всего 29)**

Напротив использование Мемов позволяет зафиксировать динамику изменения описания.

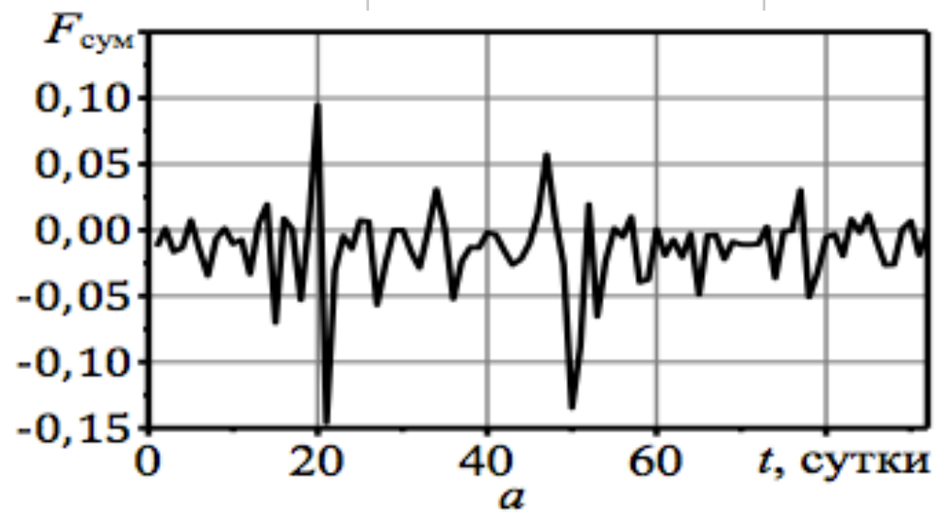
Список по ТОП15 Мемов описывающих ИП за за июнь-июль-август 2015

барак обама, боевой действие, вводить санкция, владимир путин, вооруженный сила, восток украина, гуманитарный груз, гуманитарный конвой, гуманитарный помощь, добровольческий батальон, донецкий область, зона ато, ~~июла риа, июла риа новость~~, ~~июн риа, июн риа новость~~, крушение малайзийский, крушение самолет, луганский область, местный житель, министр внутренний, народный республика, петр порошенко, президент рф, президент сша, президент украина, ~~риа новость~~, самолет сбивать, самый дело, санкционный список, санкция против, сбивать самолет, сей пора, советский союз, территория украина, тот кто, украинский военный, член экипаж **(всего 38)**

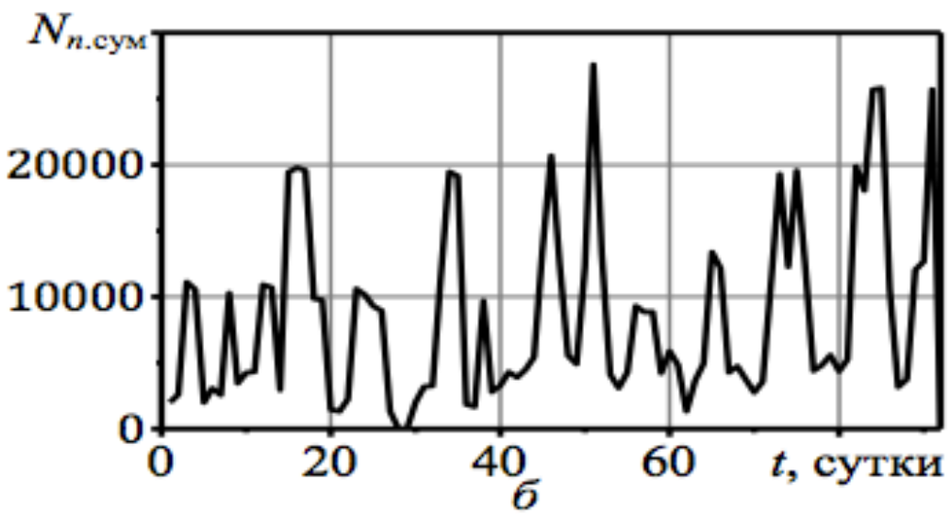
Результаты экспериментов

Так изменялось информационное пространство, измеренное через силу воздействия на мем
(**Мем-грамм модель**)

20 июня	**18 июля**	**18 августа**
луганский республика, верховный рада, сторона украина, факт обстрел, ростовский область, июн риа, сообщать риа	крушение малазийский, пассажир член экипаж, сбивать самолет, погибать пассажир, зенитноракетный комплекс	министр иностранный дело, дело украина, местный житель, ангел меркель, принимать участие, владимир владимирович,



Так изменялось информационное пространство, измеренное через число копии N-грамм
(**N-грамм модель)**



Часть 2

Статистическая математическая модель

Математическая модель

Было отобрано порядка 780 000 публикаций СМИ с 01.05.2014 г. по 01.04.2015 г., ранжируемых сервисом Meediametics.ru как наиболее цитируемые в социальных медиа. Общее количество биграмм (оригиналов по дням) в исходных данных составило 1 731 496. Анализ исходных данных показал, что изменения в ИП имеют выраженную недельную сезонную составляющую.

Гистограмма распределения числа копий биграмм по дням недели (тыс. шт.)

Математическая модель

Таблица 2 - Первые 10 наиболее популярных мемов

№	Мем	Общее количество	Среднее в день
1	«об это»	41261	156
2	«риа новость»*	36270	137
3	«владимир путин»	30932	117
4	«то число»	29200	111
5	«кроме то»	22808	86
6	«который быть»	19461	73
7	«человек который»	17336	65
8	«б б»	16616	102
9	«самый дело»	16034	61
10	«народный республика»	15273	58

Математическая модель

Таблица 2 - Группировка мемов методом дерева решении

Группа мемов	Количество копий мема в момент времени			Вероятность появления мема в период (p+1)	Число записей* (объем выборки)
	p-2	p-1	p		
1	$N(w_i^{p-2}) \geq 1$	$N(w_i^{p-1}) \geq 2$	$N(w_i^{p}) \geq 4$	0.859	457 049
2	$N(w_i^{p-2}) \geq 1$	$N(w_i^{p-1}) \geq 2$	$N(w_i^{p}) < 4$	0.596	46 894
3	$N(w_i^{p-2}) \geq 1$	$N(w_i^{p-1}) < 2$	$N(w_i^{p}) \geq 4$	0.663	372 870
4	$N(w_i^{p-2}) \geq 1$	$N(w_i^{p-1}) < 2$	$N(w_i^{p}) < 4$	0.425	143 290
5	$N(w_i^{p-2}) = 0$	$N(w_i^{p-1}) \geq 2$	$N(w_i^{p}) \geq 4$	0.507	134 646
6	$N(w_i^{p-2}) = 0$	$N(w_i^{p-1}) \geq 2$	$N(w_i^{p}) < 4$	0.271	38 146
7	$N(w_i^{p-2}) = 0$	$N(w_i^{p-1}) < 2$	$N(w_i^{p}) \geq 4$	0.417	554 664
8	$N(w_i^{p-2}) = 0$	$N(w_i^{p-1}) < 2$	$N(w_i^{p}) < 4$	0.110	557 462

Математическая модель

Будем проводить для совокупности данных, соответствующих группе 1. Для описания положительных случайных величин, аналогичных рассматриваемому нами числу копий мема в момент времени (p+1), чаще всего используются распределения из семейства экспоненциальных [8, 9]. В этом случае используются мультипликативно-экспоненциальные регрессионные уравнения следующего вида: $y = e^{BX} \cdot \varepsilon$, (5), где y – отклик; B – вектор неизвестных параметров; X – вектор входных факторов; ε – случайная ошибка.

Процедура оценивания параметров нелинейного уравнения (5) весьма затруднительна и в общем случае требует существенных алгоритмических и вычислительных затрат. Тем не менее данная проблема легко решается при переходе к логарифмам исходных данных. При логарифмировании левой и правой части уравнения (5) получаем аддитивную множественную линейную регрессионную модель $lny = BX + ln\varepsilon$, 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 $ln\left(NN(\mathrm{w}_i^{p+1})\right)$.

Математическая модель

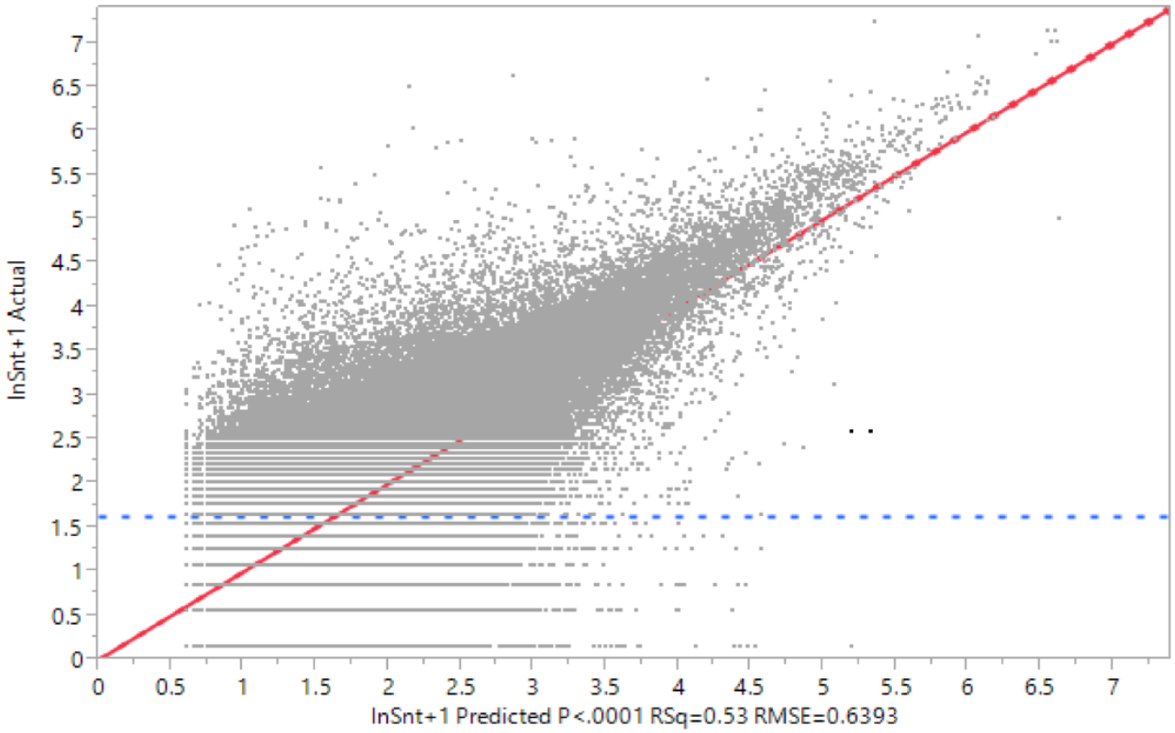
Таблица 4- Регрессоры, оказывающие статистически значимое влияние на отклик

Возможный регрессор	Коэффициент корреляции с $\boldsymbol{ln}\left(NN(\mathrm{w}_i^{p+1})\right)$	
	Обучающая выборка	Контрольная выборка
$ln\left(NN(\mathrm{w}_i^{p})\right)$	0.6692	0.6641
$ln\left(N(\mathrm{w}_i^{p-1})\right)$	0.6209	0.6248
$ln\left(N(\mathrm{w}_i^{p-2}\right)$	0.6116	0.5898
$NN(\mathrm{w}_i^{p}) \cdot NN(\mathrm{w}_i^{p-1})$	0.1485	0.1605
$NN(\mathrm{w}_i^{p-1}) \cdot NN(\mathrm{w}_i^{p-2})$	0.1434	0.1460
$\left(NN(\mathrm{w}_i^{p})\right)^2$	0.1493	0.1587
$\left(NN(\mathrm{w}_i^{p-1})\right)^2$	0.1432	0.1543
$ln\left(NN(\mathrm{w}_i^{p}\right) \cdot ln\left(NN(\mathrm{w}_i^{p-1}\right)$	0.6937	0.6965
$ln\left(NN(\mathrm{w}_i^{p-1}\right) \cdot ln\left(NN(\mathrm{w}_i^{p-2}\right)$	0.6781	0.6619
$ln\left(NN(\mathrm{w}_i^{p}\right) \cdot ln\left(NN(\mathrm{w}_i^{p-1}\right) \cdot ln\left(NN(\mathrm{w}_i^{p-2}\right)$	0.6488	0.6442
$A(\mathrm{w}_i^{p})$	0.0050	-0.0196
$F(\mathrm{w}_i^{p})$	-0.00584	-0.00037

Математическая модель

Диаграммы рассеивания

обучающая выборка



контрольная выборка

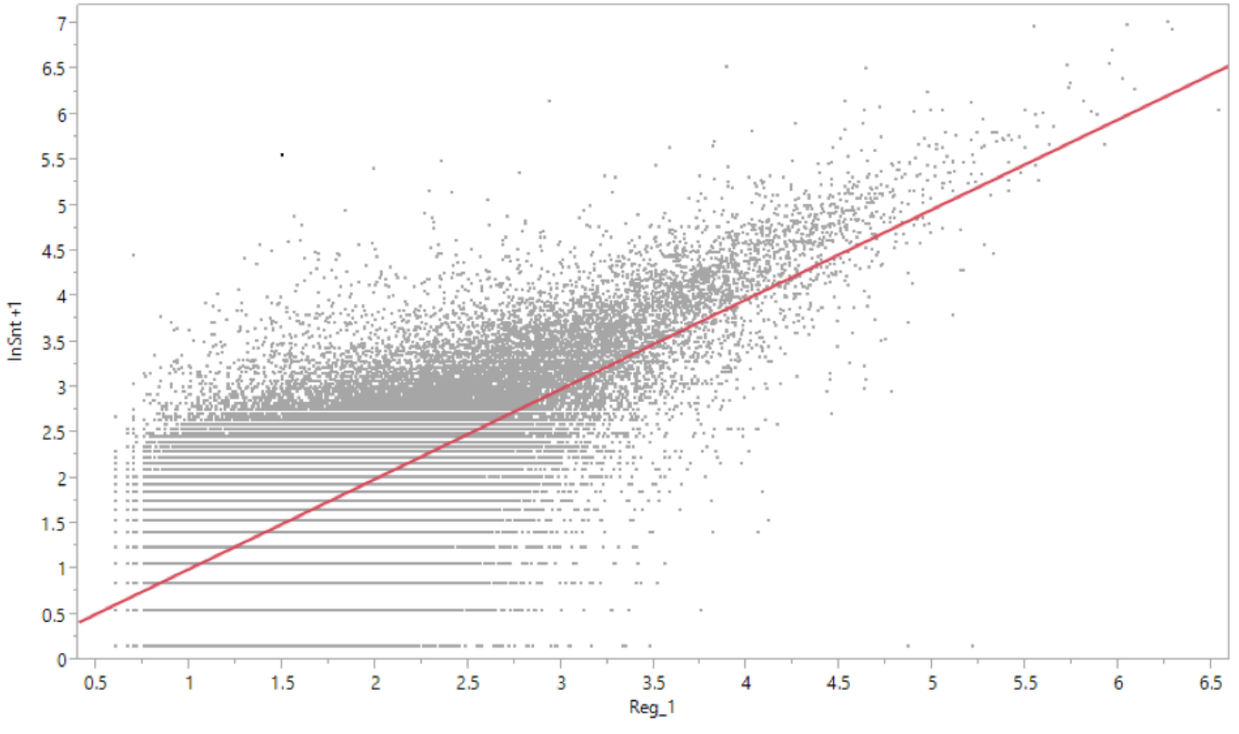


Таблица 5. - Показатели качества модели

Показатель	Обучающая выборка	Контрольная выборка
R^2	0.527452	0.516364
Средняя квадратическая ошибка	0.639258	0.648981
Число наблюдений	292 307	100 181

Математическая модель

Регрессивное уравнение модели (мат.стат модель):

$$ln\left(NN\left(\mathrm{w}_i^{p+1}\right)\right)_{\text{Прогноз}} = -0.069779 + 0.493715 \cdot ln\left(NN\left(\mathrm{w}_i^{p}\right)\right) + 0.2133316 \; ln\left(NN\left(\mathrm{w}_i^{p-1}\right)\right) + 0.2366009 \cdot ln\left(NN\left(\mathrm{w}_i^{p-2}\right)\right).$$

Отметим, что все параметры модели обладают хорошей значимостью

Таблица 6 - Статистические характеристики оценок параметров

Регрессор	Оценки параметров*	Стандартная ошибка	t-статистика	P-значения	95% доверительный интервал	
					нижняя граница	верхняя граница
константа a0	-0.069779	0.003311	-21.07	<0.0001	-0.07627	-0.06329
$ln\left(NN\left(\mathrm{w}_i^{p}\right)\right)$	0.493715	0.002325	212.35	<0.0001	0.489158	0.498272
$ln\left(NN\left(\mathrm{w}_i^{p-1}\right)\right)$	0.2133316	0.002065	103.31	<0.0001	0.209284	0.217379
$ln\left(NN\left(\mathrm{w}_i^{p-2}\right)\right)$	0.2366009	0.00179	132.2	<0.0001	0.233093	0.240109

Математическая модель

Рис.8 График динамики изменения суммы сил создания $F_t^{\text{ИП}+}$ и уничтожения $F_t^{\text{ИП}-}$

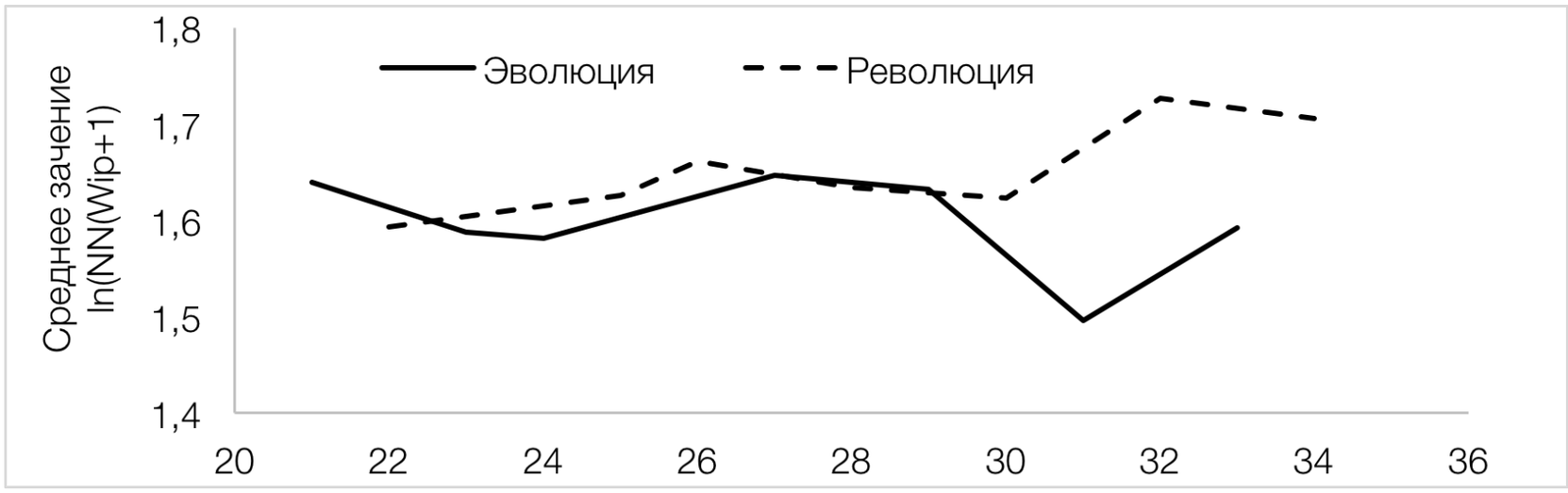


Рис. 9 График динамики изменения $ln\left(NN\left(\mathrm{w}_i^{p+1}\right)\right)$ для недель с эволюционным и революционным характером изменения ИП

Математическая модель

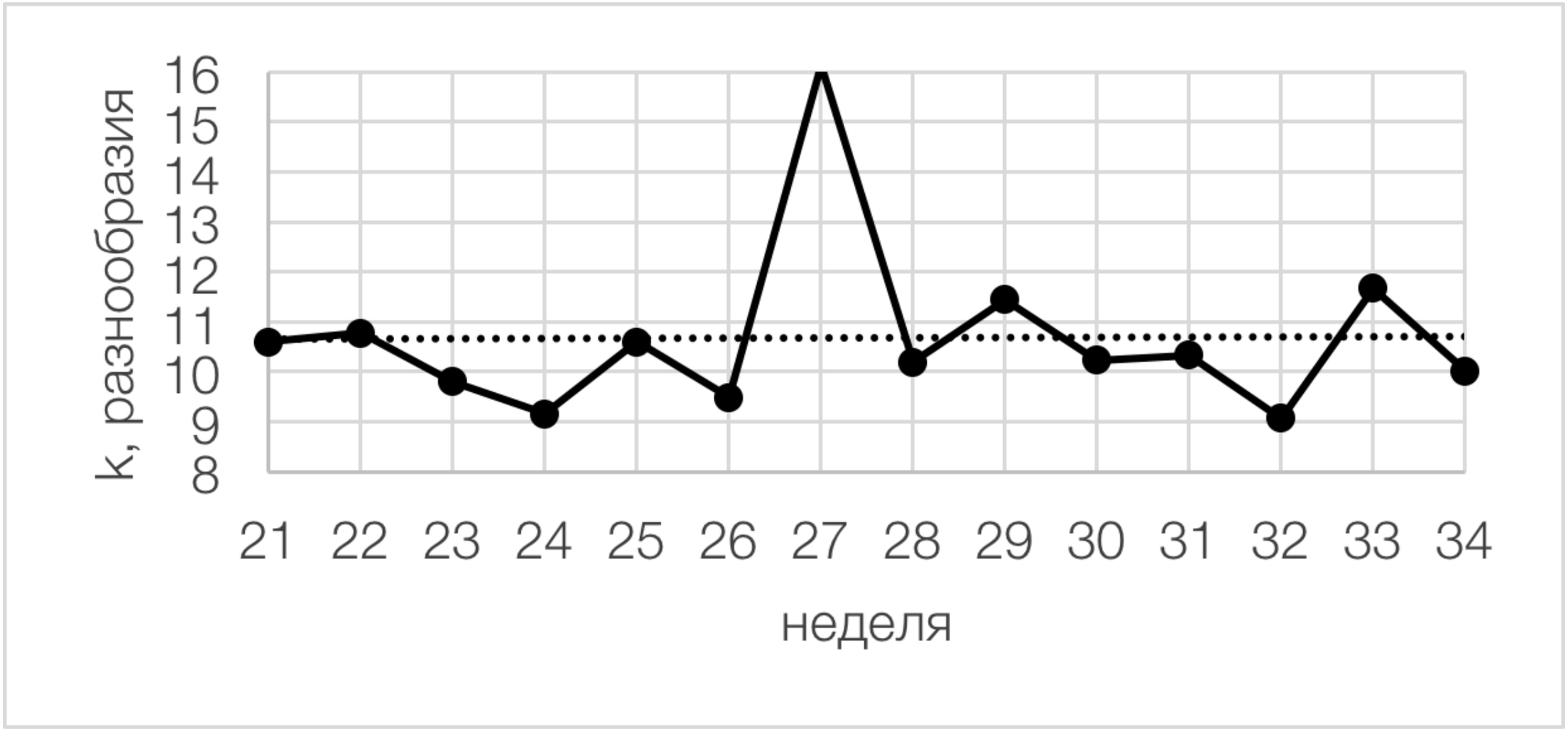


Рис. 12 График динамики $k = \frac{|C_q|}{|M_q|}$ по неделям, демонстрирующий закон сохранения разнообразия ИП

Математическая модель

Таблица 8 - Результаты 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 значения	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

*Оценка проводилась на основе метода минимального расстояния Омега-большое квадрат Мизеса. Данный метод обладает большей устойчивостью, нежели классический метод максимального правдоподобия [8, 9].

Математическая модель

Обобщим полученные результаты в виде теоретических выкладок. Рассмотрим случайную величину ξ, распределенную по закону Su-Johnson's с плотностью распределения [8,9]:

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

δ > 0 – параметр эксцесса, γ – параметр асимметрии, λ > 0 – параметр масштаба, μ – параметр сдвига, $y \in (-\infty; +\infty)$.

Основные характеристики распределения (4):

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

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

Математическая модель

Теорема 2. О плотности распределения композиции функции распределения Su-Джонсона и экспоненциальной функций, (о плотности распределения числа копии мемов).

Пусть ξ имеет плотность распределения $f_\xi(x, \delta, \gamma)$ (11), тогда случайная величина $\eta = e^\xi$ имеет плотность распределения:

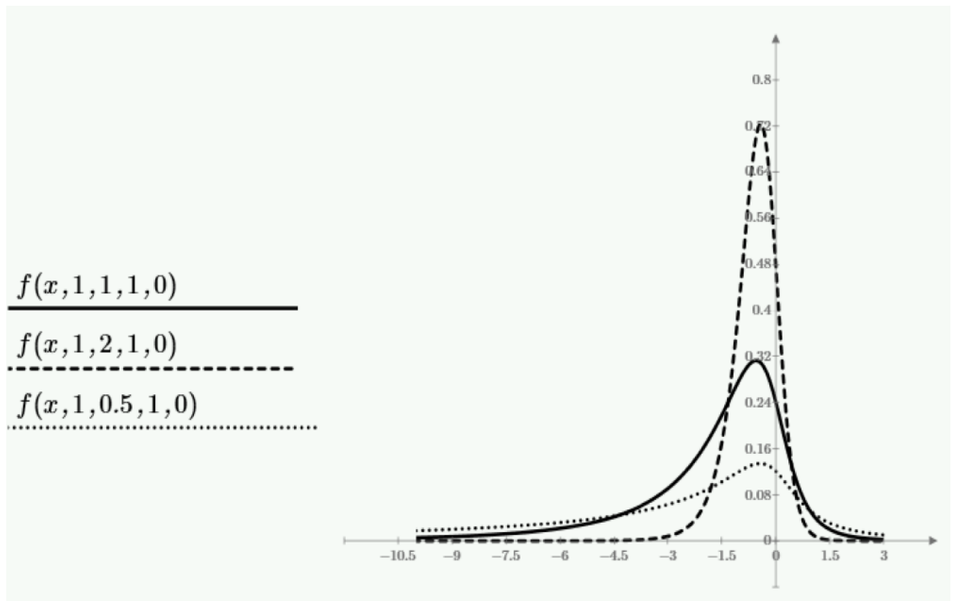
$$g(x, \gamma, \delta, \lambda, \mu) = \frac{1}{x}\frac{\delta}{\lambda\sqrt{2\pi}} \cdot \frac{1}{\sqrt{(\frac{lnx-\mu}{\lambda})^2+1}} \cdot \exp\left\{-\frac{1}{2}\left[\gamma + \delta \cdot \ln\left(\frac{lnx-\mu}{\lambda} + \sqrt{(\frac{lnx-\mu}{\lambda})^2+1}\right)\right]^2\right\}, \quad (10)$$

где $\delta > 0$ – параметр формы, γ– параметр формы, $\lambda > 0$ – параметр масштаба, μ – параметр сдвига, $y \in (-\infty; +\infty)$.

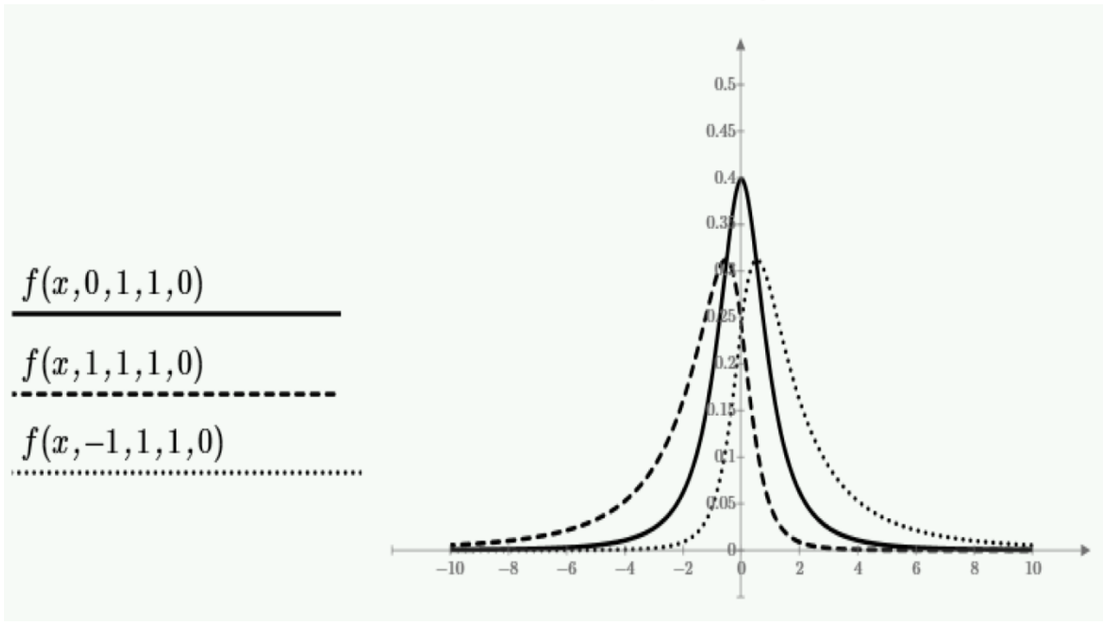
Доказательство. Сформулируем в виде вспомогательной теоремы известное из теории вероятностей утверждение о преобразовании плотности функции распределения с помощью монотонной функции [8]: пусть ξ имеет плотность распределения $f_\xi(x)$ (8) и функция $h(x)$ – монотонна, тогда случайная величина $\eta = h(\xi)$ имеет плотность распределения $f_\eta(x) = (h^{-1}(x))' \cdot f_\xi(h^{-1}(x))$. Функция $h(x) = e^x$ монотонна на всем промежутке значений x – условия вспомогательной теоремы удовлетворены, обратная функция для $h(x) = e^x$: $h^{-1}(x) = \ln(x)$. Производная от $(h^{-1}(x))' = \frac{1}{x}$. С учетом (11) и вспомогательной теоремы получаем исходное распределение (10).■

Изменения распределения Su-Джонсона в зависимости от параметров

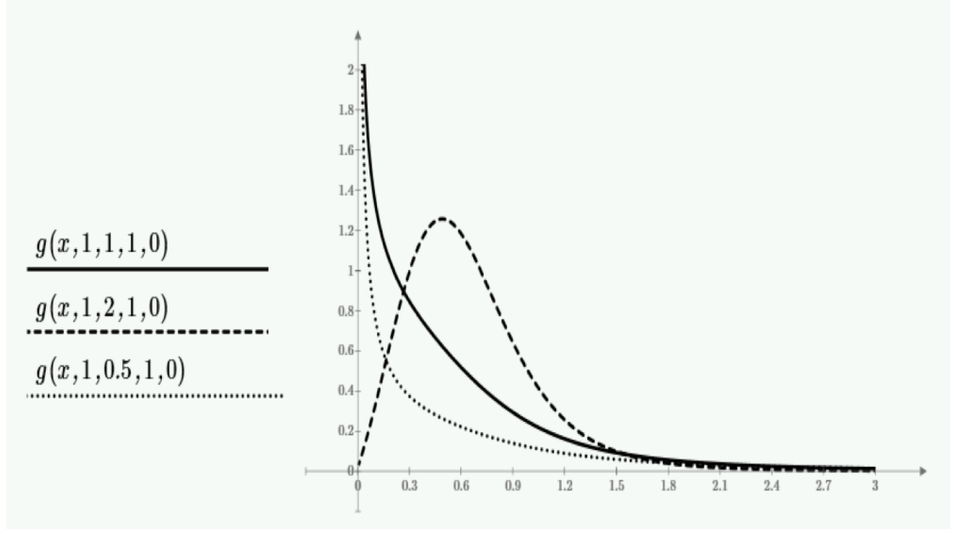
Масштаба μ



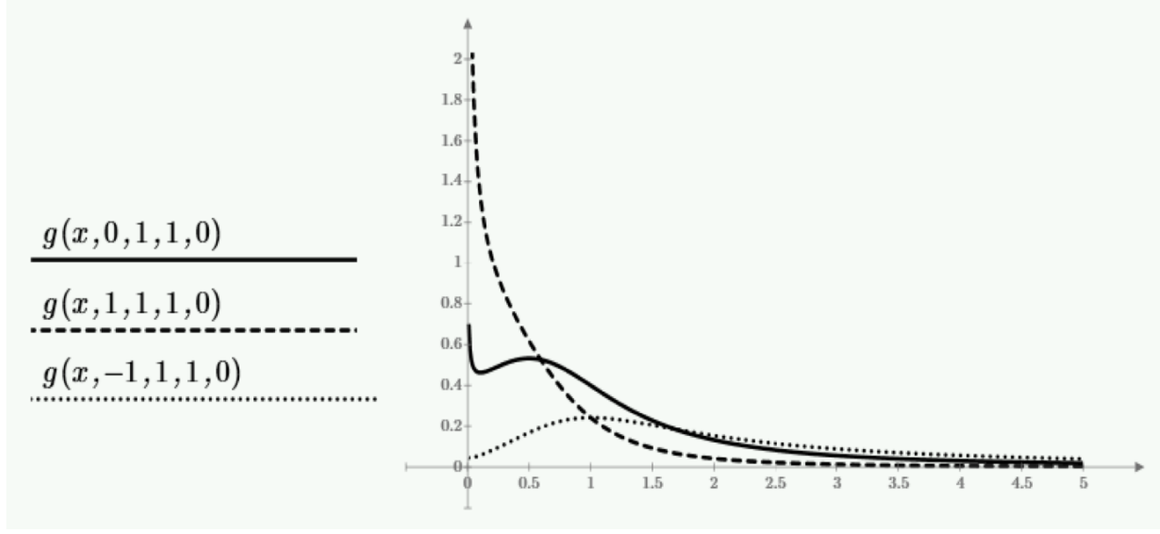
Асимметрии γ



Эксцесса δ



Сдвига λ



Изменения распределения Su-Джонсона в зависимости от параметров

Персентильные группы	Среднее значение регрессии	Асимметрия	Эксцесс	Масштаб	Сдвиг
0.005	0.62828	-2.6996	5.2407	2.6453	-0.6554
0.025	0.72407	-2.3898	4.7921	2.4874	-0.4709
0.05	0.79669	-2.0145	4.4702	2.3811	-0.238
0.1	0.87986	-1.3746	3.9188	2.2706	0.1168
0.2	0.99956	-0.727	3.4614	2.1288	0.5541
0.3	1.13747	-0.1337	3.1068	2.02	1.0491
0.4	1.26412	0.4234	2.986	1.89	1.5106
0.5	1.39745	0.7959	2.7482	1.7625	1.8948
0.6	1.54322	0.9977	2.6626	1.6367	2.1531
0.7	1.71909	1.3509	2.6217	1.4933	2.5222
0.8	1.94535	1.2469	2.4729	1.3642	2.6696
0.9	2.28319	1.1312	2.2952	1.1901	2.941
0.95	2.70246	0.9489	2.1604	1.0145	3.2495
0.975	3.11692	0.7932	1.9317	0.8624	3.602
0.995	3.69460	0.6131	1.5359	0.6571	4.1701
1	4.71040	-0.1252	0.9952	0.452	4.8353

Оценка параметров распределения Su-Джонсона

Персентильные группы	Среднее значение регрессии	Асимметрия	Эксцесс	Масштаб	Сдвиг
0.005	0.62828	-2.6996	5.2407	2.6453	-0.6554
0.025	0.72407	-2.3898	4.7921	2.4874	-0.4709
0.05	0.79669	-2.0145	4.4702	2.3811	-0.238
0.1	0.87986	-1.3746	3.9188	2.2706	0.1168
0.2	0.99956	-0.727	3.4614	2.1288	0.5541
0.3	1.13747	-0.1337	3.1068	2.02	1.0491
0.4	1.26412	0.4234	2.986	1.89	1.5106
0.5	1.39745	0.7959	2.7482	1.7625	1.8948
0.6	1.54322	0.9977	2.6626	1.6367	2.1531
0.7	1.71909	1.3509	2.6217	1.4933	2.5222
0.8	1.94535	1.2469	2.4729	1.3642	2.6696
0.9	2.28319	1.1312	2.2952	1.1901	2.941
0.95	2.70246	0.9489	2.1604	1.0145	3.2495
0.975	3.11692	0.7932	1.9317	0.8624	3.602
0.995	3.69460	0.6131	1.5359	0.6571	4.1701
1	4.71040	-0.1252	0.9952	0.452	4.8353

Оценка проводилась 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 [12]. Значения параметров подвержены существенным изменениям и явно находятся в зависимости от значений регрессионной функции

Оценка параметров распределения Su-Джонсона

В таблице 10 приведены доверительные интервалы для различных персентильных групп, построенные по данным обучающей выборки для значений для $ln\left(NN\left(\mathrm{w}_i^{p+1}\right)\right)$ и $NN\left(\mathrm{w}_i^{p+1}\right)$ с использованием распределений Su-Джонсона (8) и Теоремы1.

P (гр.)	**Среднее значение регрессии (3)**	**ДИ5% $\ln\left(\mathrm{NN}\left(\mathrm{w}_\mathrm{i}^{\mathrm{p+1}}\right)\right)$**	**ДИ95% $\ln\left(\mathrm{NN}\left(\mathrm{w}_\mathrm{i}^{\mathrm{p+1}}\right)\right)$**	**ДИ5% $\mathrm{NN}\left(\mathrm{w}_\mathrm{i}^{\mathrm{p+1}}\right)$**	**ДИ95% $\mathrm{NN}\left(\mathrm{w}_\mathrm{i}^{\mathrm{p+1}}\right)$**
0.005	0.62828	-0.123	1.798	0.884	6.037
0.025	0.72407	-0.081	1.884	0.922	6.583
0.05	0.79669	-0.045	1.93	0.956	6.891
0.2	0.99956	-0.02	2.13	0.98	8.411
0.5	1.39745	0.111	2.439	1.117	11.464
0.8	1.94535	0.685	2.888	1.984	17.963
0.975	3.11692	2.203	3.993	9.056	54.244
0.995	3.69460	2.819	4.648	16.767	104.341
1	4.71040	3.866	6.129	47.768	459.307

Оценка параметров распределения Su-Джонсона

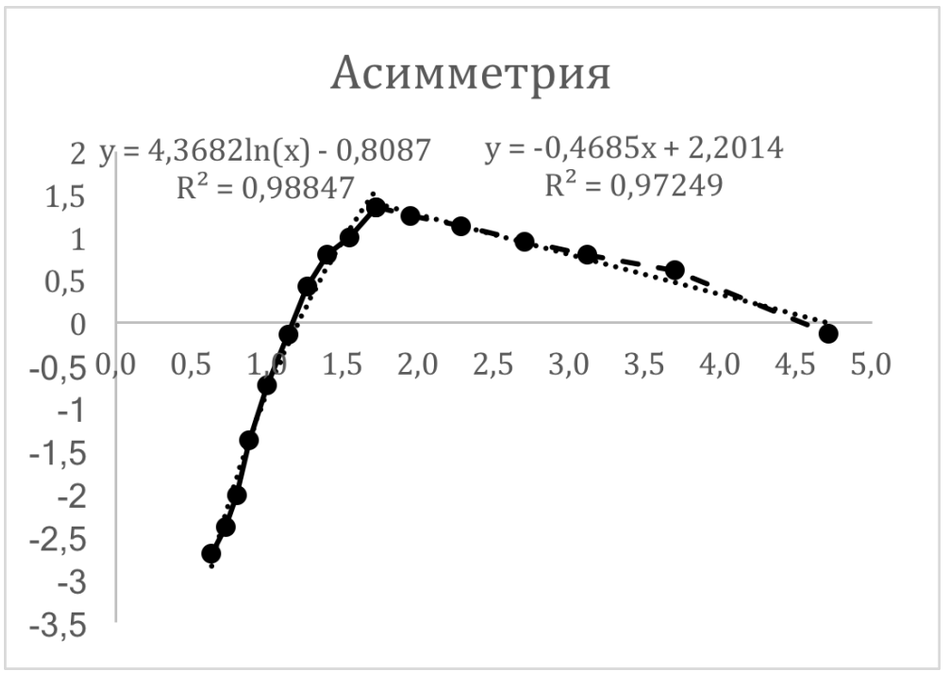
Уравнение параметров асимметрии, сдвига, эксцесса, масштаба

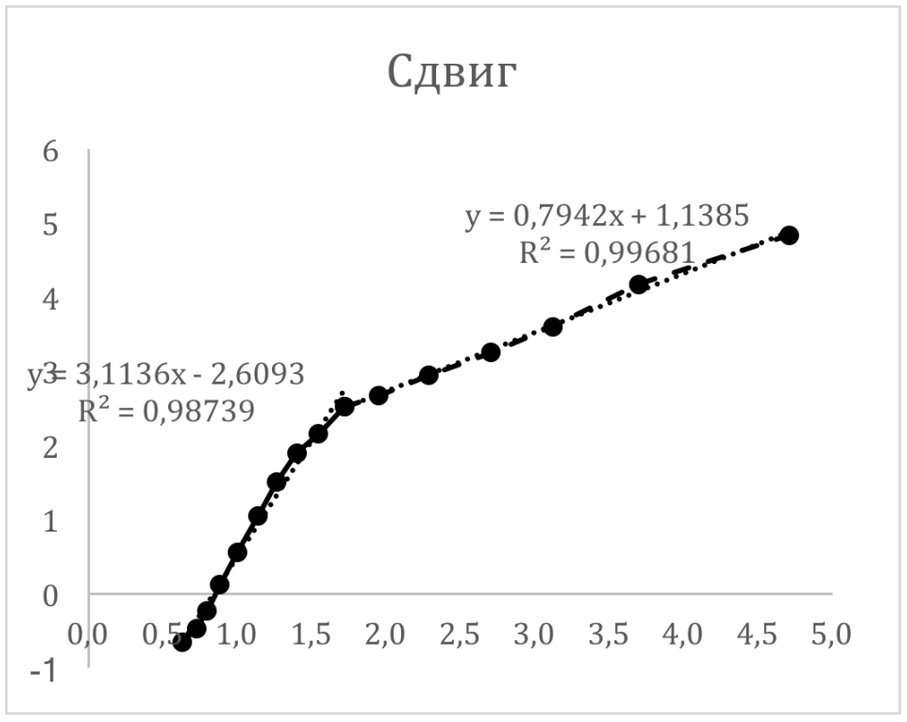
$$\gamma = \begin{cases} 3.1139x - 2{,}6093, x < 1.719 \\ 0.7942x + 1.1385, x \geq 1.719 \end{cases}$$

$$\lambda = \begin{cases} 3.0039x^2 - 9.3989x + 9.9661, x < 1.719 \\ -0.5393x + 3.5559, x \geq 1.719 \end{cases}$$

$$\delta = \begin{cases} 4.368\, ln(x) - 0.8087, x < 1.719 \\ -0.4685x + 2.2014, x \geq 1.719 \end{cases}.$$

$$\mu = -1.112\, ln(x) + 2{,}1283.$$





Оценка параметров распределения Su-Джонсона

Таблица 11 - Процент попадания значения $ln\left(NN\left(\mathbf{w}_i^{p+1}\right)\right)$ в доверительный интервал при аппроксимации Нормальным распределением

Выборка	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%
	% ошибочного попадания 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 12 - Процент попадания значения $ln\left(NN\left(\mathbf{w}_i^{p+1}\right)\right)$ в доверительный интервал при аппроксимации распределением Su-Джонсона

Выборка	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%
	% ошибочного попадания в доверительный интервал						
обучающая	1.4%	1.3%	1.4%	1.8%	0.3%	-0.6%	-0.4%
контрольная	1.0%	0.5%	0.3%	1.6%	1.5%	0.5%	0.3%

Для проверки корректности построенной модели было проведено ее тестирование на обучающей (число записей 292 307) и контрольной выборках (число записей 100 181) при условии, что все параметры модели были оценены на основе обучающей выборки.

Оценка параметров распределения Su-Джонсона

Алгоритм 1. Прогноза числа копий мема в ИП будущего момента времени

1 Вход: S_T^{IS}, $N(w_i^p)$, $N(w_i^{p-1})$, $N(w_i^{p-2})$, $Sum(N_p)$ // размер эталонного ИП, число копий мема за 3 интервала времени, размер реального ИП.

2 $k = S_T^{IS} * 10^{-6}$ // определение нормировочного коэффициента.

3 $NN(w_i^p) = \frac{N(w_i^p)}{k}$, $NN(w_i^{p-1}) = \frac{N(w_i^{p-1})}{k}$, $NN(w_i^{p-2}) = \frac{N(w_i^{p-2})}{k}$ // нормировка числа копии мема;

4 Case $(\{NN(w_i^p), NN(w_i^{p-1}), NN(w_i^{p-2})\} \Rightarrow GroupID(w_i^{p+1}) = [1 \dots 8])$ // определение группы мемов;

5 If $GroupID(w_i^{p+1}) \neq 1$ Then

6 Выход1: $N(w_i^{p+1}) = 1$; $P(w_i^{p+1}) = P(GroupID)$ // определяется вероятность, что будет хоть одна копия мема в следующий интервал времени.

7 Else

7.1 $x = -0.069779 + 0.493715 \cdot \ln\left(NN(w_i^p)\right) + 0.2133316 \cdot \ln\left(NN(w_i^{p-1})\right) + 0.2366009 \cdot \ln\left(NN(w_i^{p-2})\right)$ // прогноз логарифма числа копии.

7.2 $\gamma = \begin{cases} 3.1139x - 2{,}6093, x < 1.719 \\ 0.7942x + 1.1385, x \geq 1.719 \end{cases}$

7.3 $\delta = \begin{cases} 4.368 \ln(x) - 0.8087, x < 1.719 \\ -0.4685x + 2.2014, x \geq 1.719 \end{cases}$.

7.4 $\lambda = \begin{cases} 3.0039x^2 - 9.3989x + 9.9661, x < 1.719 \\ -0.5393x + 3.5559, x \geq 1.719 \end{cases}$

7.5 $\mu = -1.112 \ln(x) + 2{,}1283.$

7.6 $x1 = \int_0^{1000} \frac{1}{x} \frac{\delta}{\lambda\sqrt{2\pi}} \cdot \frac{1}{\sqrt{(\frac{\ln x - \mu}{\lambda})^2 + 1}} \cdot \exp\left(-\frac{1}{2}\left(\gamma + \delta \cdot \ln\left(\frac{\ln x - \mu}{\lambda} + \sqrt{(\frac{\ln x - \mu}{\lambda})^2 + 1}\right)\right)^2\right) dx = 0{,}05$

7.7 $x2 = \int_0^{1000} \frac{1}{x} \frac{\delta}{\lambda\sqrt{2\pi}} \cdot \frac{1}{\sqrt{(\frac{\ln x - \mu}{\lambda})^2 + 1}} \cdot \exp\left(-\frac{1}{2}\left(\gamma + \delta \cdot \ln\left(\frac{\ln x - \mu}{\lambda} + \sqrt{(\frac{\ln x - \mu}{\lambda})^2 + 1}\right)\right)^2\right) dx = 0{,}95$

7.8 Выход2: $P(x1 \cdot k \leq N(w_i^{p+1}) \leq x2 \cdot k) = 0{,}9$ // определение интервала числа копий мема с 90%-ой вероятностью

8 End

Оценка параметров распределения Su-Джонсона

Приведем пример использования Алгоритма1 на основе построенной статистической имитационной модели ИП для мема «15379344», представленного биграммой «июль донецкий».

Шаг 1. Задание размера пространства S_T^{IS} и определение коэффициента нормировки k в соответствии с этой размерностью.

Для описанного в отчете массива данных норма для ИП равна $S_T^{IS} = 1\,731\,496$ число копий в неделю. Для нормировки числа копий каждого мема умножаем его вес на 10^6. Таким образом, коэффициентом нормировки для числа копий мема ($Sum(N_p) = \frac{NSum(N_p) * S_T^{IS}}{10^6} = Sum(N_p) * k$) будет являться величина $k = S_T^{IS} * 10^{-6} = 1.731496$.

Оценка параметров распределения Su-Джонсона

Шаг 2. Нормировка входных данных.

Для вычисления прогноза числа копий мема в период $(p+1)$ необходимо наличие данных о числе копий этого мема за 3 предыдущих периода времени: $N(\mathrm{w}_i^p)$, $N(\mathrm{w}_i^{p-1})$, $N(\mathrm{w}_i^{p-2})$. Нормированные числа копий мема получаем делением на коэффициент нормировки: $NN(\mathrm{w}_i^p) = \frac{N(\mathrm{w}_i^p)}{k}$, $NN(\mathrm{w}_i^{p-1}) = \frac{N(\mathrm{w}_i^{p-1})}{k}$, $NN(\mathrm{w}_i^{p-2}) = \frac{N(\mathrm{w}_i^{p-2})}{k}$, расчеты приведены в таблице 17.

Таблица 13 - Исходное и нормированное количество копии и мема «15379344»

Id мема	**Неделя**	$\boldsymbol{N(\mathrm{w}_i^p)}$	$\boldsymbol{N(\mathrm{w}_i^{p-1})}$	$\boldsymbol{N(\mathrm{w}_i^{p-2})}$	$\boldsymbol{NN(\mathrm{w}_i^p)}$	$\boldsymbol{NN(\mathrm{w}_i^{p-1})}$	$\boldsymbol{NN(\mathrm{w}_i^{p-2})}$
15379344	32	15	47	98	8.663029	27.14416	56.59846

Оценка параметров распределения Su-Джонсона

Шаг 3. Определение вероятностной группы, которой принадлежит мем.

Вероятностная группировка проводится с использованием построенного дерева решений (см. Таблицу 14). Для определения группы необходимы значения $NN(\mathrm{w}_i^{p+1})$, $NN(\mathrm{w}_i^{p-1})$, $NN(\mathrm{w}_i^{p-2})$. Правила, согласно которым определяется группа, записаны в таблице 4. Для мема «15379344» по таблице 4 определяем, что он попал в группу 1.

Таблица 14 - Дерево решений для классификации группы мемов

Группа мемов	Количество копий мема момент времени			Вероятность появления мема в период (p+1)
	p-2	p-1	p	
1	$N(\mathrm{w}_i^{p-2}) \geq 1$	$N(\mathrm{w}_i^{p-1}) \geq 2$	$N(\mathrm{w}_i^{p}) \geq 4$	0.859
2	$N(\mathrm{w}_i^{p-2}) \geq 1$	$N(\mathrm{w}_i^{p-1}) \geq 2$	$N(\mathrm{w}_i^{p}) < 4$	0.596
3	$N(\mathrm{w}_i^{p-2}) \geq 1$	$N(\mathrm{w}_i^{p-1}) < 2$	$N(\mathrm{w}_i^{p}) \geq 4$	0.663
4	$N(\mathrm{w}_i^{p-2}) \geq 1$	$N(\mathrm{w}_i^{p-1}) < 2$	$N(\mathrm{w}_i^{p}) < 4$	0.425
5	$N(\mathrm{w}_i^{p-2}) = 0$	$N(\mathrm{w}_i^{p-1}) \geq 2$	$N(\mathrm{w}_i^{p}) \geq 4$	0.507
6	$N(\mathrm{w}_i^{p-2}) = 0$	$N(\mathrm{w}_i^{p-1}) \geq 2$	$N(\mathrm{w}_i^{p}) < 4$	0.271
7	$N(\mathrm{w}_i^{p-2}) = 0$	$N(\mathrm{w}_i^{p-1}) < 2$	$N(\mathrm{w}_i^{p}) \geq 4$	0.417
8	$N(\mathrm{w}_i^{p-2}) = 0$	$N(\mathrm{w}_i^{p-1}) < 2$	$N(\mathrm{w}_i^{p}) < 4$	0.110

Оценка параметров распределения Su-Джонсона

Шаг 4. Расчет прогнозного значения величины $ln\left(NN\left(\mathrm{w}_i^{p+1}\right)\right)$.

Прогнозное значение величины $ln\left(NN\left(\mathrm{w}_i^{p+1}\right)\right)$ рассчитывается с использованием регрессионной модели, которая строится для каждой из вероятностных групп отдельно. Для группы 1 значение величины $ln\left(NN\left(\mathrm{w}_i^{p+1}\right)\right)$ рассчитывается с использованием модели (7). Для мема «15379344» получаем $ln\left(NN\left(\mathrm{w}_i^{p+1}\right)\right) = 2.6941$.

Шаг 5. Расчет параметров распределения

Рассчитываем четыре параметра распределения Su-Джонсона (9) на основе значения $ln\left(NN\left(\mathrm{w}_i^{p+1}\right)\right)$ и формул (11-14). Для мема «15379344» результаты представлены в таблице 15.

Таблица 15 - Параметра распределения Su-Джонсона для логарифма количества копии мема «15379344»

Ид. мема	$\boldsymbol{ln\left(NN\left(\mathrm{w}_i^{p+1}\right)\right)}$	**Асимметрия**	**Эксцесс**	**Масштаб**	**Сдвиг**
15379344	2.6941	0.939214	2.102972	1.026237	3.278154

Оценка параметров распределения Su-Джонсона

Шаг 6. Расчет доверительных интервалов для $ln\left(NN\left(\mathrm{w}_i^{p+1}\right)\right)$ и $NN\left(\mathrm{w}_i^{p+1}\right)$.

Доверительные интервалы для $ln\left(NN\left(\mathrm{w}_i^{p+1}\right)\right)$ и $NN\left(\mathrm{w}_i^{p+1}\right)$ определяются как квантили распределений (8) и (10) соответственно. Для $ln\left(NN\left(\mathrm{w}_i^{p+1}\right)\right)$ необходимо решить численным методом интегральное уравнение для плотности (4): $\int_{-\infty}^{\infty} f(x,\gamma,\delta,\lambda,\mu)dx = p$, где p – соответствующий перцентиль (0.05 или 0.95). А для $NN\left(\mathrm{w}_i^{p+1}\right)$ - интегральное уравнение плотности (6): $\int_{0}^{\infty} g(x,\gamma,\delta,\lambda,\mu)dx = p$, где p – соответствующий перцентиль (0.05 или 0.95). При численном решении интегральных уравнений следует вместо знака ∞ подставлять любое заведомо большое число, для данных данного исследования достаточно подставить число 1000. В расчетах используются параметры распределения, полученные на шаге 5. Для мема «15379344» получим результаты, представленные в таблице 16.

Таблица 16 - Параметры распределения Su-Джонсона для количества копий мема «15379344»

Ид. мема	$ln\left(NN\left(\mathrm{w}_i^{p+1}\right)\right)$	$NN\left(\mathrm{w}_i^{p+1}\right)$	**ДИ5%** $ln\left(NN\left(\mathrm{w}_i^{p+1}\right)\right)$	**ДИ95%** $ln\left(NN\left(\mathrm{w}_i^{p+1}\right)\right)$	**ДИ5%** $NN\left(\mathrm{w}_i^{p+1}\right)$	**ДИ95%** $NN\left(\mathrm{w}_i^{p+1}\right)$
15379344	1.935921	6.352888	1.666	3.632	5.296	37.797

Оценка параметров распределения Su-Джонсона

Графики распределений, полученных для мема «15379344

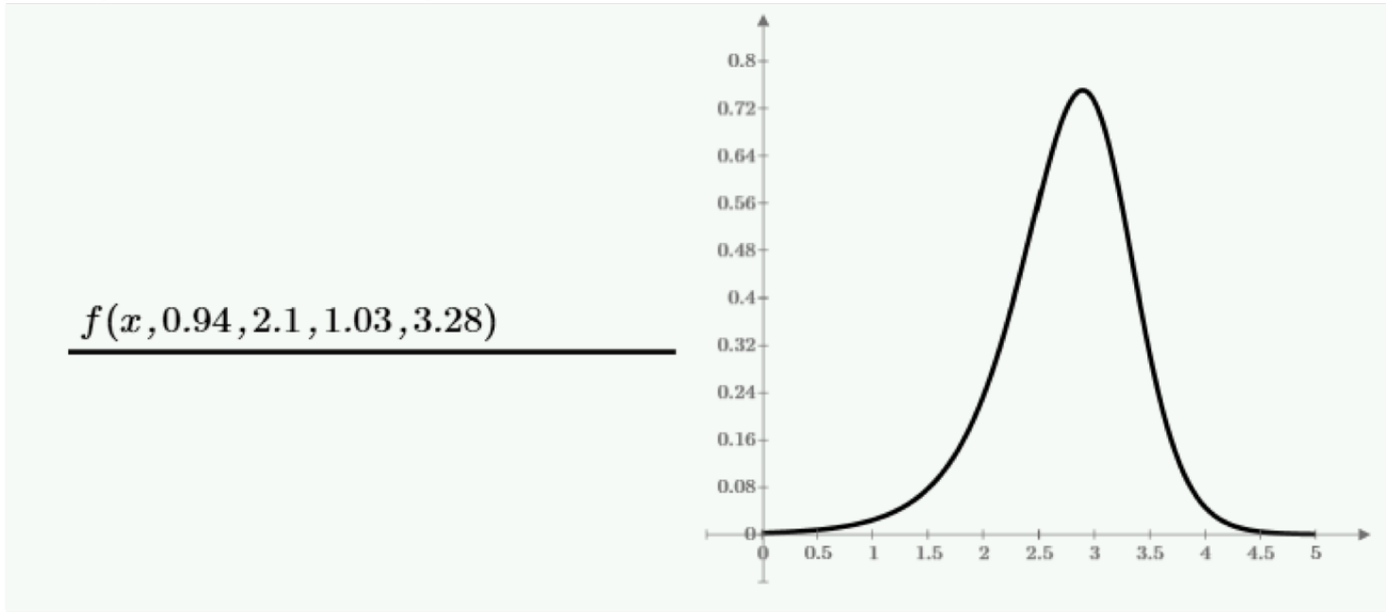


Рис. 45 График $f(x, \gamma, \delta, \lambda, \mu)$ для $ln\left(NN\left(\mathrm{w}_i^{p+1}\right)\right)$ мема 15379344

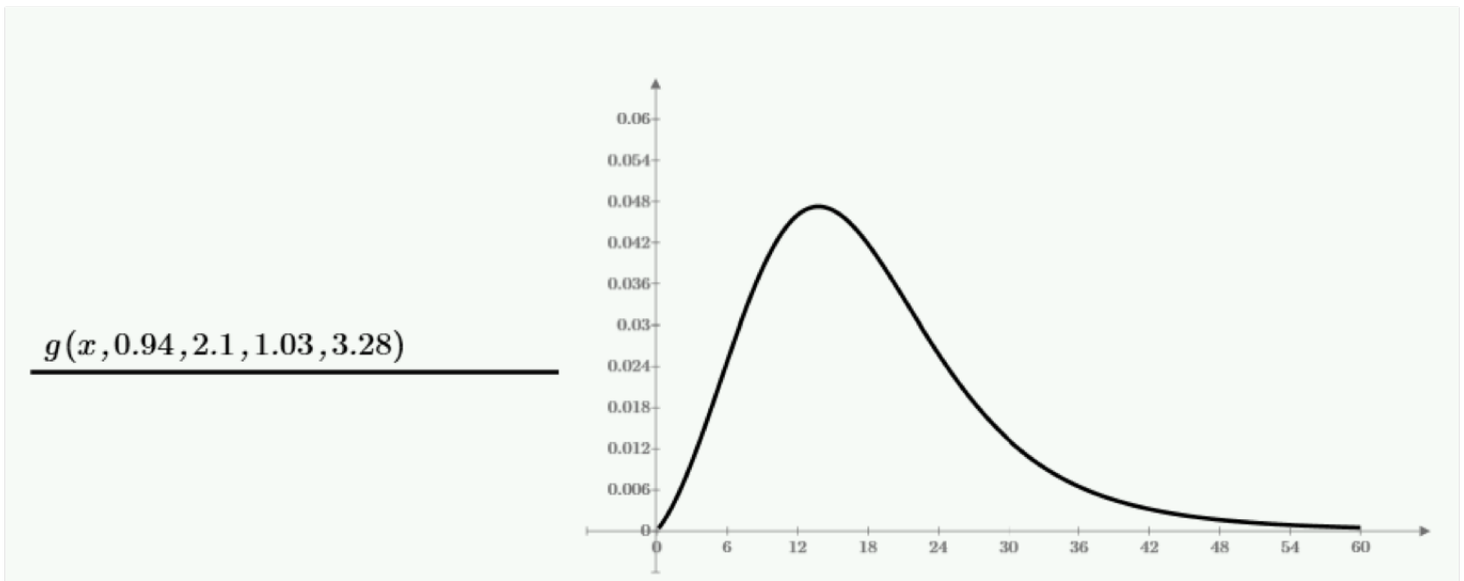


Рис. 46 График $g(x, \gamma, \delta, \lambda, \mu)$ для $NN\left(\mathrm{w}_i^{p+1}\right)$ мема 15379344

Оценка параметров распределения Su-Джонсона

Шаг 7. Расчет доверительных интервалов для $NN(\mathrm{w}_i^{p+1})$.

Для обратного перехода от нормированных величин к исходным необходимо провести ренормализацию, то есть умножить значения границ доверительных интервалов ДИ5% $NN(\mathrm{w}_i^{p+1})$ и ДИ95% $Ng(\mathrm{w}_i^{p+1})$ на коэффициент нормировки k, определенный на шаге 1. Для мема «15379344» получаем результаты, представленные в таблице 17.

Таблица 17 - 90% - доверительный интервал для количества копии мема «15379344»

Ид. мема	$\boldsymbol{NN(\mathrm{w}_i^{p+1})}$ ***(реальное значение)***	**ДИ5%** $\boldsymbol{NN(\mathrm{w}_i^{p+1})}$	**ДИ95%** $\boldsymbol{NN(\mathrm{w}_i^{p+1})}$
15379344	11	9.17	65.45

Таким образом, был получен 90% доверительный интервал прогноза для мема «15379344» в период 33-й недели. Ожидаемый доверительный интервал находится в промежутке от 9.17 до 65.45 числа копий. Как видим, реальное значение числа копий в 33-ю неделю составляет 11 копий, что попадает в наш прогнозный интервал.

Эпилог

Эволюция

Алгоритм 1. Прогноза числа копий мема в ИП будущего момента времени

1 Вход: S_T^{15}, $N(w_i^p)$, $N(w_i^{p-1})$, $N(w_i^{p-2})$, $Sum(N_p)$ // размер эталонного ИП, число копий мема за 3 интервала времени, размер реального ИП.

2 $k = S_T^{15} * 10^{-6}$ // определение нормировочного коэффициента.

3 $NN(w_i^p) = \frac{N(w_i^p)}{k}$, $NN(w_i^{p-1}) = \frac{N(w_i^{p-1})}{k}$, $NN(w_i^{p-2}) = \frac{N(w_i^{p-2})}{k}$ // нормировка числа копии мема;

4 Case $(\{NN(w_i^p), NN(w_i^{p-1}), NN(w_i^{p-2})\} \Rightarrow GroupID(w_i^{p+1}) = \{1 \dots 8\})$ // определение группы мемов;

5 If $GroupID(w_i^{p+1}) \neq 1$ Then

6 Выход1: $N(w_i^{p+1}) = 1$; $P(w_i^{p+1}) = P(GroupID)$ // определяется вероятность, что будет хоть одна копия мема в следующий интервал времени.

7 Else

7.1 $x = -0.069779 + 0.493715 \cdot \ln(NN(w_i^p)) + 0.2133316 \cdot \ln(NN(w_i^{p-1})) + 0.2366009 \cdot \ln(NN(w_i^{p-2}))$ // прогноз логарифма числа копии.

7.2 $\gamma = \begin{cases} 3.1139x - 2{,}6093, x < 1.719 \\ 0.7942x + 1.1385, x \geq 1.719 \end{cases}$

7.3 $\delta = \begin{cases} 4.368\ln(x) - 0.8087, x < 1.719 \\ -0.4685x + 2.2014, x \geq 1.719 \end{cases}$

7.4 $\lambda = \begin{cases} 3.0039x^2 - 9.3989x + 9.9661, x < 1.719 \\ -0.5393x + 3.5559, x \geq 1.719 \end{cases}$

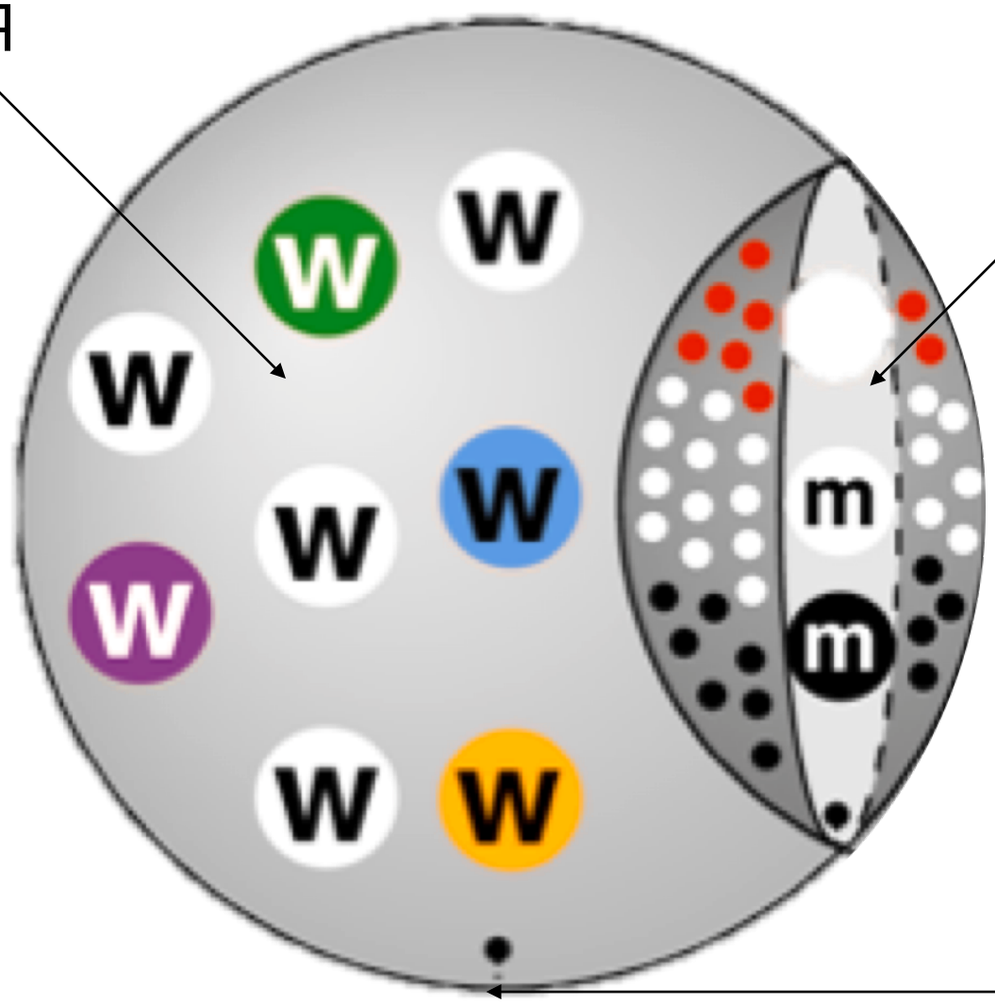
7.5 $\mu = -1.112\ln(x) + 2{,}1283.$

7.6 $x1 = \int_0^{1000} \frac{1}{x} \frac{\delta}{\lambda\sqrt{2\pi}} \cdot \frac{1}{\sqrt{\left(\frac{\ln x - \mu}{\lambda}\right)^2 + 1}} \cdot \exp\left(-\frac{1}{2}\left(\gamma + \delta \cdot \ln\left(\frac{\ln x - \mu}{\lambda} + \sqrt{\left(\frac{\ln x - \mu}{\lambda}\right)^2 + 1}\right)\right)^2\right) dx = 0{,}05$

7.7 $x2 = \int_0^{1000} \frac{1}{x} \frac{\delta}{\lambda\sqrt{2\pi}} \cdot \frac{1}{\sqrt{\left(\frac{\ln x - \mu}{\lambda}\right)^2 + 1}} \cdot \exp\left(-\frac{1}{2}\left(\gamma + \delta \cdot \ln\left(\frac{\ln x - \mu}{\lambda} + \sqrt{\left(\frac{\ln x - \mu}{\lambda}\right)^2 + 1}\right)\right)^2\right) dx = 0{,}95$

7.8 Выход2: $P(x1 \cdot k \leq N(w_i^{p+1}) \leq x2 \cdot k) = 0{,}9$ // определение интервала числа копий мема с 90%-ой вероятностью

8 End



Число

$$\frac{|C_q|}{|M_q|} \approx \alpha$$

Целое **W***

слова ***мем*** *и слова* ***мемплекс*** *и слова*

*В начале было Слово, и **Слово было** у Бога, и **Слово было Бог***

*первая строка Евангелия от Иоанна

Спасибо за внимание!