УДК 519.688+004.942

Исследование статистических характеристик текста на основе графовой модели лингвистического корпуса

Е. Г. Григорьева, В. А. Клячин

Григорьева Елена Геннадиевна, кандидат физико-математических наук, доцент кафедры компьютерных наук и экспериментальной математики, Волгоградский государственный университет, Россия, 400062, г. Волгоград, Университетский пр-т, д. 100, e_grigoreva@volsu.ru

Клячин Владимир Александрович, доктор физико-математических наук, заведующий кафедрой компьютерных наук и экспериментальной математики, Волгоградский государственный университет, Россия, 400062, г. Волгоград, Университетский пр-т, д. 100; Калмыцкий государственный университет имени Б. Б. Городовикова, Россия, Республика Калмыкия, 358000, г. Элиста, ул. Пушкина, д. 11 klyachin.va@volsu.ru

Статья посвящена исследованию статистических характеристик текста, которые вычисляются на базе графовой модели представления текста из лингвистического корпуса. Во введении излагается актуальность статистического анализа текстов и приводятся некоторые задачи, решаемые с помощью такого анализа. Предлагаемая в статье графовая модель текста строится как граф, в вершинах которого расположены слова текста, а ребра графа отражают факт попадания двух слов в какую-либо часть текста, например в предложение. Для вершин и ребер графа в статье вводятся понятия веса как значения из некоторой аддитивной полугруппы. Доказываются формулы вычисления графа и его весов при конкатенации текстов. На основе предложенной модели реализуются вычисления на языке программирования Python. Для экспериментального исследования статистических характеристик выделяются 24 величины, которые выражаются через веса вершин, ребер графа, а также других характеристик графа, например степени его вершин. Надо отметить, что целью численных экспериментов является поиск характеристик текста, с помощью которых можно определять, является ли текст созданным человеком или случайно сгенерированным. В статье предлагается один из возможных таких алгоритмов, который генерирует случайный текст, используя некоторый созданный человеком другой текст в качестве шаблона. При этом в случайном тексте сохраняется последовательность чередования частей речи вспомогательного текста. Оказывается, что требуемым условиям удовлетворяет медианное значение отношения величины веса ребра графа текста к числу предложений в тексте.

Ключевые слова: текст, лингвистический корпус, граф, автоматическая обработка текста.

Поступила в редакцию: 28.02.2019 / Принята: 19.05.2019 / Опубликована: 02.03.2020

DOI: https://doi.org/10.18500/1816-9791-2020-20-1-116-126

ВВЕДЕНИЕ

В работе предложена модель лингвистического корпуса на основе графа отношений принадлежности слов семантическим или синтаксическим единицам

текста. Из весовых значений вершин и ребер этого графа могут быть получены статистические данные, позволяющие решать различные актуальные задачи лингвистики: извлечение ключевых слов, оценка сочетаемости слов, тематическое моделирование лингвистических корпусов, построение конкордансов и т. д. А также предложенная модель может быть взята за основу вычисления признаков текстов для применения в исследованиях, опирающихся на методы машинного обучения. Приведены алгебраические свойства графовой модели, позволяющие вычислять весовые параметры вершин и ребер графа при склеивании (конкатенации) текстов или объединении текстов в единый корпус. Предложенная модель, в отличие от часто используемой модели «мешка слов», позволяет использовать определенную структуру текста, закладываемую в модель в зависимости от способа его разбиения, при применении статистических методов анализа. Использование таких методов достаточно распространено и применяется при решении ряда задач компьютерной корпусной лингвистики. Например, в работе [1] описывается процесс автоматической обработки текстового корпуса, собранного из новостных лент ряда интернет-сайтов, для создания вероятностной n-граммной модели разговорного русского языка. Приводится статистический анализ данного корпуса, даются результаты по подсчету частоты появления различных n-грамм слов. В работе [2] статистическими методами в рамках модели «мешка слов» с применением наивного байесовского классификатора исследуется возможность ранжирования русскоязычных текстов по их эмоциональной тональности в соответствии с классификацией Г. Левхейма. В работе [3] рассматриваются алгоритмы определения семантической близости ключевых слов: алгоритм Гинзбурга, основанный на частотных характеристиках слов, и его программная реализация, а также алгоритм с учетом частей речи и проблемы его реализации. В работе [4] предлагается рассматривать статистические параметры текста в качестве авторской характеристики, используя различные информационные и программные средства. Отметим также работы [5, 6], в которых предлагаются методы машинного обучения и соответствующие модели векторного представления лингвистических текстов для решения различных задач компьютерной лингвистики.

Отметим, что описываемая в настоящей статье графовая модель обобщает упомянутую выше модель «мешка слов» и содержит необходимую информацию для анализа текста, к примеру, методами законов Зипфа. Напомним, что классические законы Зипфа используют статистику отдельных слов текста для определения его осмысленности. В первом законе Зипфа утверждается, что ранг частоты слова (т. е. номер частоты в отсортированном по убыванию массиве частот) убывает обратно пропорционально рангу. Второй закон Зипфа утверждает, что количество слов с заданной частотой экспоненциально убывает относительно частоты в упорядоченном по возрастанию массиве частот. Экспериментально проверено выполнение этих законов для текстов, написанных человеком, — осмысленных текстов. Однако не сложно предложить конструкцию, которая позволяет сгенерировать бессмысленный текст с выполнением указанных законов. Коротко опишем эту конструкцию. По заданному тексту создается словарь слов и вычисляются их частоты. Далее моделируется случайная величина (номер слова в словаре), имеющая такое же распределение, как и слова в заданном тексте. Соответствующий дискретный случайный процесс будет генерировать случайный текст, статистически близкий к тексту, полученному перемешиванием слов исходного текста. При перемешивании осмысленность текста исчезнет. Таким образом, законы Зипфа в рамках модели «мешка слов» не позволяют отличить тексты, которые могут быть сгенерированы указанным способом.

В настоящей статье предложена графовая модель текста и дано экспериментальное подтверждение того, что в рамках этой модели могут быть вычислены такие характеристики исходного текста, которые способны отличать данные тексты от текстов, которые генерируются не только вышеуказанным способом, но и рядом других способов, описываемых ниже.

1. ОПИСАНИЕ ГРАФОВОЙ МОДЕЛИ

Алфавитом будем называть произвольное конечное множество Σ. Элементы множества Σ будем называть символами. Упорядоченный набор символов назовем словом или цепочкой символов. Множество всех цепочек символов алфавита Σ обозначим через Σ^*. Текстом будем называть упорядоченный набор символов и записывать его в виде

$$T = a_1 a_2 ... a_n, \quad a_i \in \Sigma, \quad n = |T|.$$

Здесь и далее через $|X|$ обозначим мощность множества X — количество элементов множества X. Если T_1, T_2 — два текста, то через $T_1 \cdot T_2$ будем обозначать текст, склеенный из двух данных текстов (конкатенация текстов). Через 2^X будем обозначать множество всех подмножеств множества X. Некоторую совокупность текстов будем называть *корпусом*. Для построения графовой модели нам необходимо понятие разбиения текста, под которым будем понимать произвольное подмножество $P \subset 2^N$ упорядоченных подмножеств множества $N = \{1, 2, 3, ..., |T|\}$. Каждое такое подмножество определяется набором кортежей номеров $(i_1, ..., i_k) \in P$ входящих в него символов. Примерами таких разбиений могут быть разбиение текста на слова, разбиение текста на предложения и разбиение текста на m-граммы.

Пусть $I \in P$. Построим отображение $\omega : P \to \Sigma^*$

$$\omega(I) = a_{i_1} a_{i_2} ... a_{i_k}, \quad i_1 < i_2 < ... < i_k, \quad i_j \in I, \quad j = 1, 2, ..., k.$$

Это отображение сопоставляет набору номеров символов соответствующую часть текста в виде цепочки символов.

Пример 1. Зафиксируем произвольный символ $s \in \Sigma$. Для заданного текста T пусть найдутся номера $1 = i_0, < i_1 < ... < i_k < i_{k+1} = |T|$ такие, что $a_{i_j} = s$. Тогда

$$P_s = [(i_0, ..., i_1 - 1), (i_1 + 1, ..., i_2 - 1), ..., (i_k + 1, ..., |T|)]$$

представляет собой разбиение текста с помощью разделителя $s \in \Sigma$.

Пример 2. Для заданного разбиения вида P_s и произвольного натурального m можно построить разбиение вида

$$P_{s,m} = \omega_1 \cup ... \omega_m \cup \omega_2 \cup ... \omega_{m+1} \cup ... \omega_{k+m-2} \cup ... \omega_k,$$

где

$$\omega_j = (i_{j-1} + 1, ..., i_j - 1), \quad j = 1, ..., k, \quad \omega_{k+1} = (i_k + 1, ..., i_{k+1}).$$

Данное разбиение представляет собой разбиение m-грамм, а словами служат части ω_j исходного разбиения.

По определению будем считать, что разбиение P' является более мелким, чем разбиение P'', и записывать $P' \subset P''$, если для всякого $I' \in P'$ найдется такое $I'' \in P''$, что $I' \subset I''$. Для двух заданных текстов T_1, T_2 с соответствующими разбиениями P_1, P_2 определим разбиение $P = P_1 \circ P_2$ для текста $T_1 \cdot T_2$ следующим образом:

$$P_1 \circ P_2 = \{I \subset (1, 2, ..., |T_1| + |T_2|) : \ I = I_1 \subset P_1 \text{ или } I = \{i + |T_1| : \ i \in I_2 \in P_2\}\}.$$

Теорема 1. *Совокупность всех текстов и их разбиений* (T, P) *образует полугруппу с групповой операцией*

$$(T_1, P_1) + (T_2, P_2) = (T_1 \cdot T_2, P_1 \circ P_2).$$

Доказательство. Единственное, что требуется проверить, — это ассоциативность введенной выше операции. Рассмотрим три текста с какими-либо разбиениями (T_i, P_i), $i = 0, 1, 2$. Ясно, что

$$(T_1 \cdot T_2) \cdot T_3 = T_1 \cdot (T_2 \cdot T_3).$$

Обозначим $P_2 \circ P_3 = P_2'$, $T_2 \cdot T_3 = T_2'$. По определению имеем

$$P_1 \circ (P_2 \circ P_3) = \{I \subset (1, 2, ..., |T_1| + |T_2'|) : \ I = I_1 \subset P_1,$$
$$\text{или } I = \{i + |T_1| : \ i \in I_2', \ I_2' \in P_2'\}\} = \{I \subset (1, 2, ..., |T_1| + |T_2| + |T_3|) : \ I = I_1 \subset P_1,$$
$$\text{или } I = \{i + |T_1| : \ i \in I_2, \ I_2 \in P_2, \ \text{или } i \in \{j + |T_2|, \ j \in I_3, \ I_3 \in P_3\}\}\}.$$

Так, элементы $I \in P_1 \circ (P_2 \circ P_3)$ имеют один из трех видов

$$I \in P_1, \quad I = \{i + |T_1|, i \in I_2, \ I_2 \in P_2\}, \quad I = \{j + |T_1| + |T_2|, j \in I_3, \ I_3 \in P_3\}.$$

Не сложно проверить, что мы получим тот же результат для $I \in (P_1 \circ P_2) \circ P_3$. Так, окончательно

$$(T_1, P_1) + ((T_2, P_2) + (T_3, P_3)) = ((T_1, P_1) + (T_2, P_2)) + (T_3, P_3).$$

Что и требовалось доказать. □

Для заданного текста T и его разбиения P обозначим через $U(T, P) = \{\omega(I) : I \in P\}$ совокупность уникальных частей разбиения. Например, в случае разбиения текста на слова это множество представляет собой совокупность уникальных слов текста. Рассмотрим текст T и два его произвольных разбиения $P' \subset P''$. Построим граф $G = G(T, P', P'')$ с множеством вершин $VG = U(T, P')$ и множеством ребер

$$EG = \{(a, b), a, b \in U(T, P') : \ \exists \ I \in P'', \ I_a, I_b \subset I, \ \text{где } \omega(I_a) = a, \ \omega(I_b) = b\}.$$

Этот граф представляет отношение принадлежности пары частей текста одного разбиения одной части другого разбиения. Например, ребру графа с вершинами в виде уникальных слов текста соответствует пара слов, которые встречаются в каком-либо одном предложении. Непосредственно можно проверить следующую ключевую формулу для построения графа при склеивании двух текстов.

Теорема 2. *Пусть имеются тексты* T_1, T_2 *с соответствующими разбиениями* $P_i' \subset P_i''$, $i = 1, 2$. *Тогда имеет место равенство*

$$G(T_1 \cdot T_2, P_1' \circ P_2', P_1'' \circ P_2'') = G(T_1, P_1', P_1'') \cup G(T_2, P_2', P_2''). \tag{1}$$

Доказательство. Заметим, что $VG(T_1 \cdot T_2, P_1' \circ P_2', P_1'' \circ P_2'') = U(T_1 \cdot T_2, P_1 \circ P_2) = U(T_1, P_1) \cup U(T_2, P_2) = VG(T_1, P_1) \cup VG(T_2, P_2)$. Далее, если имеется ребро из $e \in EG(T_1 \cdot T_2, P_1' \circ P_2', P_1'' \circ P_2'')$, то существует пара частей w_1, w_2 из $P_1' \circ P_2'$, входящая в $P_1'' \circ P_2''$. Предположим, что ребро (w_1, w_2) не принадлежит ни $EG(T_1, P_1', P_1'')$, ни $EG(T_2, P_2', P_2'')$. Это может быть, только если они не попали ни в одну из частей P_1''

и ни в одну из частей P_2''. Поскольку при конкатенации текстов части P_1'', P_2'' объединяются, то пара не попадает ни в одну часть из $P_1'' \circ P_2''$. Получили противоречие. Теорема доказана. □

Формула (1) позволяет строить граф G объединения двух и более текстов целого лингвистического корпуса для последующего анализа статистических величин, связанных с весами вершин и ребер этого графа. Мы рассмотрим наиболее общий случай весовых значений. Это позволит выполнить обобщенную реализацию вычисления весов при конкатенации текстов.

Рассмотрим некоторую аддитивную полугруппу W. Вес для графа w — это отображение, сопоставляющее каждой вершине графа или каждому ребру графа значения из полугруппы W

$$w:\ VG \to W \quad \text{или} \quad w:\ EG \to W.$$

Введение весовой функции для построенного графа объясняется тем, что в корпусной лингвистике распространены статистические методы решения разнообразных задач. Весовая функция и призвана производить вычисление соответствующих статистических величин. Для вершин графа $G(T, P', P'')$ определим следующее множество:

$$B(v) = \{J \in P'':\ \exists\ \alpha, \beta \in \Sigma^*,\ \omega''(J) = \alpha v \beta\}.$$

Другими словами, $B(v)$ — подмножество разбиения P'', которому соответствуют части текста, содержащие цепочку v.

Пример 3. Пусть $W = \mathbb{N}$ множество натуральных чисел. Для $v \in VG$ и $G = G(T, P', P'')$ положим $w'(v) = |\omega'^{-1}(v)|$, $w''(v) = |B(v)|$.

Первая функция вычисляет количество вхождений уникальной части текста v разбиения P' в тексте T, а вторая функция вычисляет количество частей текста разбиения P'', содержащих v. Пусть $\mathbb{Z}_m$ обозначает циклическую группу порядка $m \in \mathbb{N}$. На множестве подмножеств множества $\mathbb{Z}_m$ введем операцию «+» следующим образом:

$$I_1 + I_2 = sort(I_1 \cup I_2), \quad I_1, I_2 \in 2^{\mathbb{Z}_m},$$

как результат композиции объединения и сортировки. Положим $m = |P''|$. Упорядочим каким-нибудь образом множество P''. Тогда каждому $J \in P''$ можно сопоставить некоторый номер $n(J) \in \mathbb{Z}_m$, а множеству $B(v)$ — некоторое подмножество $I \subset \mathbb{Z}_m$. Введенная выше операция делает множество $2^{\mathbb{Z}_m}$ полугруппой. Тогда можно положить

$$w^m(v) = B(v) \in 2^{\mathbb{Z}_m}.$$

Данная весовая функция строит упорядоченные списки номеров частей текста разбиения P'', содержащие уникальную часть текста v разбиения P'.

Определив некоторые весовые функции, которые вычисляют конкретные характеристики текста, необходимо научиться преобразовывать веса при операции склеивания текстов. Для двух текстов T_1, T_2 и соответствующих весовых функций $w_i:\ VG(T_i, P_i', P_i'') \to W$ определим весовую функцию

$$w(v) = \begin{cases} w_1(v), & \text{если } v \in U(T_1, P_1') \setminus U(T_2, P_2'), \\ w_2(v), & \text{если } v \in U(T_2, P_2') \setminus U(T_1, P_1'), \\ w_1(v) + w_2(v), & \text{если } v \in U(T_2, P_2') \cap U(T_1, P_1'), \end{cases}$$

Эта формула позволяет вычислять весовую функцию для результата склейки двух текстов при условии, что значения весовых функций двух текстов принадлежат одному множеству. Получим также формулу для случая, когда эти множества разные. Пусть m_1, m_2 — два натуральных числа. Ясно, что $2^{\mathbb{Z}_{m_1}} \subset 2^{\mathbb{Z}_{m_1+m_2}}$. Пусть $I_1 \in 2^{\mathbb{Z}_{m_1}}$, $I_2 \in 2^{\mathbb{Z}_{m_2}}$. Положим

$$I_1 \oplus I_2 = \{(i_1, ..., i_k, j_{1+m_1}, ..., j_{l+m_1}) : \ (i_1, ..., i_k) \in I_1, \ (j_1, ..., j_l) \in I_2\}.$$

Непосредственным вычислением проверяется свойство ассоциативности

$$(I_1 \oplus I_2) \oplus I_3 = I_1 \oplus (I_2 \oplus I_3).$$

Приведем пример весовой функции, которая может иметь определенное значение при исследовании лингвистического корпуса. Предположим, что имеются два текста T_i, $i = 1, 2$ с некоторыми разбиениями $P'_i \subset P''_i$, $i = 1, 2$ и две весовые функции $w_i : \ VG(T_i, P'_i, P''_i) \to W_i$, где $W_i = 2^{\mathbb{Z}_{m_i}}$, где $m_i = |P''_i|$. Требуемую функцию построим следующим образом:

$$w(v) = \begin{cases} w_1(v), & \text{если } v \in U(T_1, P'_1) \setminus U(T_2, P'_2), \\ \varnothing \oplus w_2(v), & \text{если } v \in U(T_2, P'_2) \setminus U(T_1, P'_1), \\ w_1(v) \oplus w_2(v), & \text{если } v \in U(T_2, P'_2) \cap U(T_1, P'_1). \end{cases}$$

Приведем теперь пример весовой функции для ребер графа $G = G(T, P', P'')$. Обозначим $m = |P''|$. Пусть $e = (v_1, v_2) \in EG$. Примерами весовых функций могут быть функции

$$w(e) = |w^m(v_1) \cap w^m(v_2)|, \quad \text{или} \quad w(e) = w^m(v_1) \cap w^m(v_2).$$

Первая функция вычисляет количество вхождений пары (v_1, v_2) в одну часть разбиения P'', вторая функция перечисляет все такие части разбиения P''. Заметим, что эти функции могут быть вычислены через соответствующие весовые функции вершин графа.

2. ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ ГРАФОВОЙ МОДЕЛИ

Графовая модель текста была реализована на языке программирования Python. Реализация включает в себя четыре основных класса: TEXT, PARTITION, STATISTICS, GRAPH. Первые два представляют текст так таковой и его различные разбиения. Класс GRAPH хранит структуру графа текста по его разбиениям $P' \subset P''$. Наконец, класс STATISTICS вычисляет статистические характеристики текста на построенном графе. В экспериментах использовались разбиения P' — на слова, P'' — на предложения. Анализ проводился на базе текстов художественной литературы русских классиков.

Все эксперименты посвящены исследованию характеристик текста, по которым возможно отличать текст, написанный человеком, от текста, случайно построенного на том же наборе слов, что и исходный текст. Ниже мы приводим общую схему генерации такого текста.

Рассмотрим произвольный текст и извлечем из него слова, из которых составим словарь. В предыдущем разделе этот словарь обозначается через $U(T, P')$. Для каждого элемента словаря вычислим частоту появления соответствующего слова в тексте. Эту частоту будем принимать в качестве вероятности слова в тексте. Рассмотрим несколько моделей случайного текста.

В первой модели в словарь добавим специальное слово — символ разделителя предложений. После этого на каждой итерации процесса построения текста разыгрывается случайная величина, распределение которой совпадает с распределением слов исходного текста. При этом значением данной случайной величины можно считать номер слова в словаре. По этому случайному номеру находится слово и добавляется в текст. Заметим, что статистически этот процесс эквивалентен простому случайному перемешиванию слов и разделителей предложений исходного текста.

Во второй модели мы также добавим символ разделителя предложений. Каждому слову в словаре добавим некоторый его морфологический признак. Это может быть, например, часть речи — существительное, прилагательное, глагол, наречие и т. д. Процесс построения нового текста читает исходный текст и заменяет каждое слово случайным словом с тем же морфологическим признаком, что и текущее слово. Заметим, что эта модель статистически эквивалентна случайному перемешиванию слов исходного текста, в котором перемешивание осуществляется по слоям морфологических признаков. При этом сохраняется структура каждого предложения текста: последовательность морфологических признаков слов предложения сохраняется. Ясно, что эта вторая модель случайного текста является частным случаем первой модели.

Введем следующие обозначения. Через $L(v)$, $v \in VG$ обозначим последовательность значений весов ребер, инцидентных вершине v, через $\deg(v)$ обозначим степень вершины v, через $\theta(e)$ обозначим вес ребра $e \in EG$, наконец, через $\sigma(T)$ обозначим число предложений текста T — или, другими словами, число элементов разбиения P''. Далее для всякой конечной последовательности $x = \{x_1, x_2, ..., x_n\}$ введем обозначения $\max(x)$, $\mathrm{mean}(x)$, $\mathrm{median}(x)$ для максимального, среднего и медианного значения в x. Для текста T обозначим соответствующие последовательности:

$$d(T) = \{d(v), v \in VG\}, \quad \theta(T) = \{\theta(e), e \in EG\},$$
$$\theta_s(T) = \{\theta(e)/\sigma(T), e \in EG\}, \quad d_{mx}(T) = \{\max(L(v)), v \in VG\},$$
$$d_{mn}(T) = \{\mathrm{mean}(L(v)), v \in VG\}, \quad d_{mdn}(T) = \{\mathrm{median}(L(v)), v \in VG\}.$$

Отметим, что распределение величины степеней вершин случайного графа [7–9] является предметом теоретического исследования и имеет практическое применение, например, при анализе графовой структуры сети Интернет. При этом предполагается, что степени имеют плотность распределение Парето $p(x) = x^{-t}$ [10, 11].

В качестве числовых признаков текстов для исследования мы рассматриваем следующие характеристики графа $G(T, P', P'')$:

– среднее значение степени вершин графа, $\mathrm{mean}(d(T))$;
– максимальное значение степени вершин графа, $\max(d(T))$;
– медианное значение степени вершин графа, $\mathrm{median}(d(T))$;
– среднее значение весов ребер графа, $\mathrm{mean}(\theta(T))$;
– максимальное значение весов ребер графа, $\max(\theta(T))$;
– медианное значение весов ребер графа, $\mathrm{median}(\theta(T))$.

Кроме этих величин будем также рассматривать следующие характеристики:

$$\max(\theta_s(T)), \quad \mathrm{mean}(\theta_s(T)), \quad \mathrm{median}(\theta_s(T)), \quad \mathrm{std}(\theta_s(T)),$$
$$\max(d_{mx}(T)), \quad \mathrm{mean}(d_{mx}(T)), \quad \mathrm{median}(d_{mx}(T)), \quad \mathrm{std}(d_{mx}(T)),$$
$$\max(d_{mn}(T)), \quad \mathrm{mean}(d_{mn}(T)), \quad \mathrm{median}(d_{mn}(T)), \quad \mathrm{std}(d_{mn}(T)),$$

$$\max(d_{mdn}(T)), \quad \mathrm{mean}(d_{mdn}(T)), \quad \mathrm{median}(d_{mdn}(T)), \quad \mathrm{std}(d_{mdn}(T)).$$

Здесь $\mathrm{std}(x)$ обозначает стандартное отклонение в последовательности x

$$\mathrm{std}(x) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x - \mathrm{mean}(x))^2}.$$

Определенный интерес могут представлять величины, для которых значения на исходном тексте и на случайном тексте отличаются, например, в не менее 70% всех текстов. Из представления результатов на рис. 1 видно, что этому условию будут удовлетворять величины

$$\mathrm{mean}(\theta_s(T)), \quad \mathrm{median}(\theta_s(T)), \quad \mathrm{std}(\theta(T)), \quad \mathrm{mean}(\theta(T)).$$

При этом первые две величины дают требуемое отличие в 84 % случаев. Другими словами, в 84 % текстов эти величины для исходных текстов меньше, чем для случайных текстов, сгенерированных по исходным алгоритмам, приведенным выше.

Теперь рассмотрим поведение тех же величин, если в качестве исходного текста выбирается случайный текст, построенный по исходному. Те же вычисления дают результат, приведенный на рис. 2.

Из этого графика видно, что только величина $\mathrm{median}(\theta_s(T))$ дает нужные результаты, поскольку показывает, что случайно сгенерированные исходные тексты в большинстве случаев (97%) неразличимы с текстами, которые случайным образом сгенерированы по ним же. Таким образом, медианное значение величины веса ребра графа $G(T, P', P'')$, деленного на количество предложений в тексте T, может быть использовано в качестве величины, определяющей, является ли исходный текст осмысленным или он сгенерирован случайно по некоторому шаблонному тексту, используя алгоритм, описанный выше.

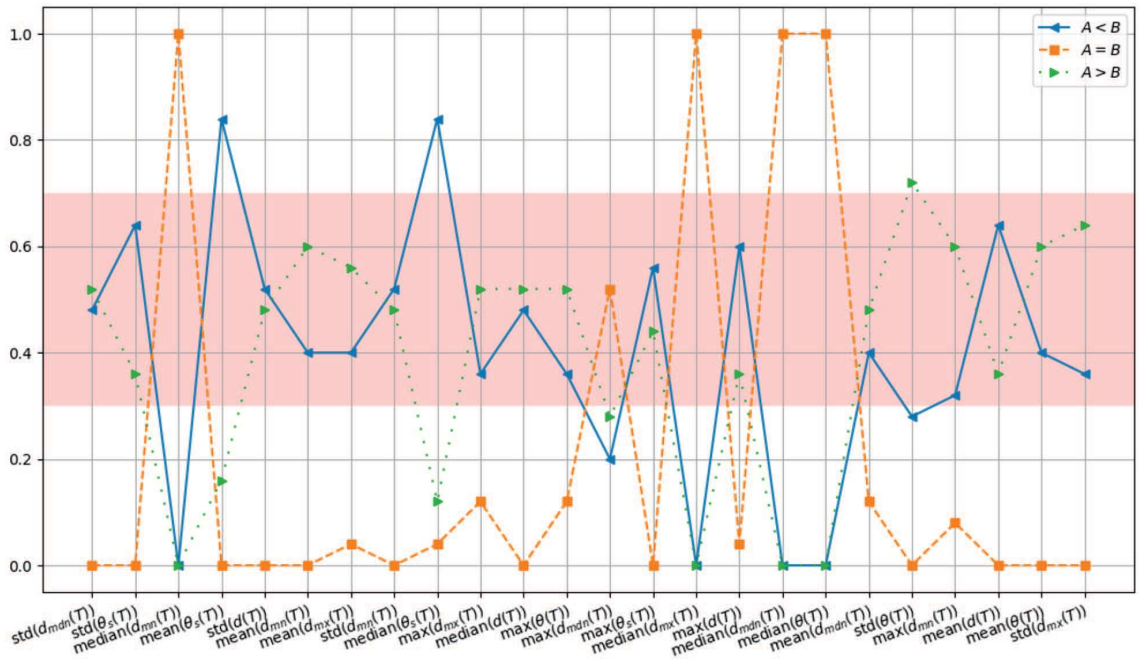

Рис. 1. Результаты численного анализа характеристик текста: A — список значений для исходного текста; B — список значений для случайного текста (цвет online)

Fig. 1. Results of numerical analysis of text characteristics: A — list of values for source text; B — list of values for random generated text (color online)

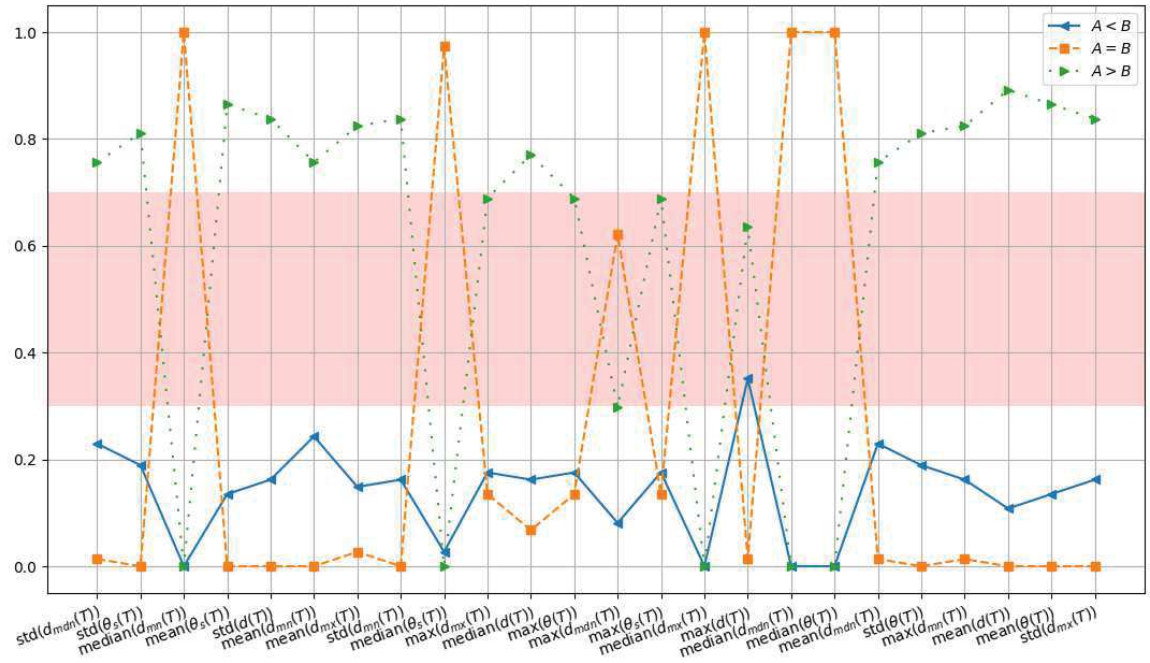

Рис. 2. Результаты численного анализа характеристик текста после второго эксперимента: A — список значений для исходного текста; B — список значений для случайного текста (цвет online)

Fig. 2. Results of numerical analysis of text characteristics after the second experiment: A — list of values for source text; B — list of values for random generated text (color online)

Благодарности. Работа выполнена при финансовой поддержке РФФИ и Администрации Волгоградской области (проект № 18-412-340007).

Библиографический список

1. *Кипяткова И. С., Карпов А. А.* Автоматическая обработка и статистический анализ новостного текстового корпуса для модели языка системы распознавания русской речи // Информационно-управляющие системы. 2010. № 4 (47). С. 2–8.
2. *Колмогорова А. В., Калинин А. А., Маликова А. В.* Лингвистические принципы и методы компьютерной лингвистики для решения задач сентимент-анализа русскоязычных текстов // Актуальные проблемы филологии и педагогической лингвистики. 2018. № 1 (29). С. 139–148. DOI: https://doi.org/10.29025/2079-6021-2018-1(29)-139-148
3. *Воронина И. Е., Кретов А. А., Попова И. В.* Алгоритмы определения семантической близости ключевых слов по их окружению в тексте // Вестн. ВГУ. Сер. Системный анализ и информационные технологии. 2010. № 1. С. 148–153.
4. *Берман Н. Д., Левенец А. В., Сергеева Л. А.* Статистический анализ текстовой информации // Информационные технологии XXI века : сб. науч. тр. / отв. за вып. Е. А. Шеленок. Хабаровск : Изд-во Тихоокеан. гос. ун-та, 2016. С. 282–286.
5. *Донина О. В.* Применение методов Data Mining для решения лингвистических задач // Вестн. ВГУ. Сер. Системный анализ и информационные технологии. 2017. № 1. С. 154–160.
6. *Mikolov T., Chen K., Corrado G., Dean J.* Efficient Estimation of Word Representations in Vector Space. arxiv.org/abs/1301.3781v3
7. *Райгородский А. М.* Случайные графы // Математика в задачах. Сборник материалов выездных школ команды Москвы на Всероссийскую математическую олимпиаду / под ред. А. А. Заславского, Д. А. Пермякова, А. Б. Скопенкова, М. Б. Скопенкова, А. В. Шаповалова. М. : Изд-во МЦНМО, 2009. С. 312–315.

8. *Erdős P., Rányi A.* On random graphs I // Publ. Math. Debrecen. 1959. Vol. 6. P. 290–297.
9. *Newman M. E. J., Strogatz S. H., Watts D. J.* Random graphs with arbitrary degree distribution and their applications // Phys. Rev. E. 2001. Vol. 64. P. 26–118.
10. *Павлов Ю. Л., Чеплюкова И. А.* Случайные графы Интернет-типа и обобщенная схема размещения // Дискрет. матем. 2008. Т. 20, вып. 3. С. 3–18. DOI: https://doi.org/10.4213/dm1008
11. *Павлов Ю. Л.* О предельных распределениях степеней вершин в условных Интернет-графах // Дискрет. матем. 2009. Т. 21, вып. 3. С. 14–23. DOI: https://doi.org/10.4213/dm1057

Образец для цитирования:
Григорьева Е. Г., Клячин В. А. Исследование статистических характеристик текста на основе графовой модели лингвистического корпуса // Изв. Сарат. ун-та. Нов. сер. Сер. Математика. Механика. Информатика. 2020. Т. 20, вып. 1. С. 116–126. DOI: https://doi.org/10.18500/1816-9791-2020-20-1-116-126

The Study of the Statistical Characteristics of the Text Based on the Graph Model of the Linguistic Corpus

E. G. Grigorieva, V. A. Klyachin

Elena G. Grigorieva, https://orcid.org/0000-0001-8303-262X, Volgograd State University, 100 Universitetskii Prosp., Volgograd 400062, Russia, e_grigoreva@volsu.ru
Vladimir A. Klyachin, https://orcid.org/0000-0003-1922-7847, Volgograd State University, 100 Universitetskii Prosp, Volgograd 400062, Russia; Kalmyk State University name after B. B. Gorodovikov, 11 Pushkin St., Elista 358000, Repablic of Kalmykia, Russia, klyachin.va@volsu.ru

The article is devoted to the study of the statistical characteristics of the text, which are calculated on the basis of the graph model of the text from the linguistic corpus. The introduction describes the relevance of the statistical analysis of the texts and some of the tasks solved using such an analysis. The graph model of the text proposed in the article is constructed as a graph in the vertices of which the words of the text are located, and the edges of the graph reflect the fact that two words fall into any part of the text, for example, in — a sentence. For the vertices and edges of the graph, the article introduces the concept of weight as a value from some additive semigroup. Formulas for calculating a graph and its weights are proved for text concatenation. Based on the proposed model, calculations are implemented in the Python programming language. For an experimental study of statistical characteristics, 24 values are distinguished, which are expressed in terms of the weights of the vertices, edges of the graph, as well as other characteristics of the graph, for example, the degrees of its vertices. It should be noted that the purpose of numerical experiments is to squeak in the characteristics of the text, with which you can determine whether the text is man-made or randomly generated. The article proposes one of the possible such algorithms, which generates random text using some other text created by man as a template. In this case, the sequence of parts of speech in an auxiliary text alternation is preserved in the random text. It turns out that the required conditions are satisfied by the median value of the ratio of the text graph edge weight value to the number of sentences in the text.

Keywords: text, graph, linguistic corpus, automatic text processing.

Received: 28.02.2019 / Accepted: 19.05.2019 / Published: 02.03.2020

Acknowledgements: This work was supported by the Russian Foundation for Basic Research and the Administration of the Volgograd Region (project No. 18-412-340007).

References

1. Kipyatkova I. S., Karpov A. A. Automatic processing and statistic analysis of the news text corpus for a language model of a Russian language speech recognition system. *Informatsionno-upravliayuschie sistemy* [Information and Control Systems], 2010, no. 4 (47), pp. 2–8 (in Russian).
2. Kolmogorova A. V., Kalinin A. A., Malikova A. V. Linguistic principles and computational linguistics methods for the purposes of sentiment analysis of Russian texts. *Aktual'nye problemy filologii i pedagogicheskoi lingvistiki* [Actual problems of philology and pedagogical linguistics], 2018, no. 1 (29), pp. 139–148 (in Russian). DOI: https://doi.org/10.29025/2079-6021-2018-1(29)-139-148
3. Voronina I. E., Kretov A. A., Popova I. V. Algorithms of semantic proximity assessment based on the lexical environment of the key words in a text. *Proceedings of Voronezh State University. Ser. Systems analysis and information technologies*, 2010, no. 1, pp. 148–153 (in Russian).
4. Berman N. D., Levenets A. V., Sergeeva L. A. Statistical analysis of textual information. In: *Informatsionnye tekhnologii XXI veka* [Information Technologies of the XXI Century. Collection of Scientific Papers]. Khabarovsk, Izdatel'stvo Tikhookeanskgo gosudarstvennogo universiteta, 2016, pp. 282–286 (in Russian).
5. Donina O. V The application of data mining methods in linguistics. *Proceedings of Voronezh State University. Ser. Systems analysis and information technologies*, 2017, no. 1, pp. 154–160 (in Russian).
6. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. arxiv.org/abs/1301.3781v3
7. Raigorodskii A. M. Random Graphs. In: *Matematika v zadachakh* [Mathematics in Problems]. Moscow, Izdatel'stvo Moskovskogo tsentra nepreryvnogo matematicheskogo obrazovaniya, 2009, pp. 312–315 (in Russian).
8. Erdős P., Rányi A. On random graphs I. *Publ. Math. Debrecen*, 1959, vol. 6, pp. 290–297.
9. Newman M. E. J., Strogatz S. H., Watts D. J., Random graphs with arbitrary degree distribution and their applications. *Phys. Rev. E*, 2001, vol. 64, pp. 26–118.
10. Pavlov Yu. L., Cheplyukova I. A. Random graphs of Internet type and the generalised allocation scheme. *Discrete Mathematics and Applications*, 2008, vol. 18, iss. 5, pp. 447–463. DOI: https://doi.org/10.1515/DMA.2008.033
11. Pavlov Yu. L. On the limit distributions of the vertex degrees of conditional Internet graphs. *Discrete Mathematics and Applications*, 2009, vol. 19, iss. 4, pp. 349–359. DOI: https://doi.org/10.1515/DMA.2009.023

Cite this article as:
Grigorieva E. G., Klyachin V. A. The Study of the Statistical Characteristics of the Text Based on the Graph Model of the Linguistic Corpus. *Izv. Saratov Univ. (N. S.), Ser. Math. Mech. Inform.*, 2020, vol. 20, iss. 1, pp. 116–126 (in Russian). DOI: https://doi.org/10.18500/1816-9791-2020-20-1-116-126