РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ

А В Т О М Е Т Р И Я

2005, том 41, № 4

МОДЕЛИРОВАНИЕ В ФИЗИКО-ТЕХНИЧЕСКИХ ИССЛЕДОВАНИЯХ

УДК 519.24

Е. Л. Кулешов, В. В. Крысанов, К. Какушо

(Владивосток, Россия – Кусацу, Киото, Япония)

РАСПРЕДЕЛЕНИЕ ВЕРОЯТНОСТЕЙ ЧАСТОТЫ СЛОВА В ТЕКСТАХ*

Предлагается новая математическая модель формирования распределения вероятностей частоты слова в текстах, таких как английский, русский, английский гипертекст. Получено распределение вероятностей частот, обобщающее закон Парето. Представлен алгоритм вычисления параметров модели. Показано, что полученное распределение вероятностей обеспечивает существенно более высокую степень согласия с экспериментальными данными, чем распределение Парето.

Введение. Частота использования слова (число использований слова) в том или ином документе, тексте является важнейшей характеристикой, влияющей на динамику связи в современной Интернет-сети. Например, такие системы поиска информации, как Google, Yahoo, Altavista и т. п., используют алгоритмы, основанные на вычислении функции разрешающей способности (значимости) слова относительно текста, которая непосредственно определяется распределением вероятностей случайной величины [1–3]. Использование более точной модели формирования распределения вероятностей частоты позволяет повысить эффективность алгоритмов автоматической классификации текстов [4], а также алгоритмов поиска цифровых документов [5, 6].

Принято считать, что – дискретная случайная величина с распределением вероятностей $P(\quad k) \quad b \,/ k \quad ^{1}$, где $b \quad 0, \quad 0, \; k \quad 1,2,\ldots$ и $P(\quad k)$ вероятность случайного события k. Это соотношение известно в литературе как распределение Парето [7], а также как второй закон Ципфа [8] – важнейший эмпирический закон, позволяющий моделировать сложные недетерминированные системы (например, телекоммуникационые сети [9]) и некоторые стохастические процессы (например, количество ролей, сыгранных актерами Голливуда в художественных фильмах [9], рост численности персонала компаний [10] и т. д.). В работах [11, 12] авторы выделяют три основ-

* Работа выполнена при финансовой поддержке Японского агентства науки и технологий (проект «Универсальный дизайн цифрового города»).

ных механизма, предположительно порождающих данное распределение вероятностей: 1) как выбор, основанный на предпочтениях (механизм впервые предложен в [13]), 2) как следствие оптимизационных процессов [14], а также 3) как следствие мультипликативных стохастических процессов различной природы [15]. В работе Саймона [13] показано, что частота слова в тексте, а также численность населения городов имеют распределение Парето. Аналогично Мандельброт в работе [14] дает теоретическое обоснование первого закона Ципфа: $q_n \quad dn$, где n – ранг (порядковый номер) слова в последовательности слов, упорядоченных по убыванию частоты использования (наименьший ранг у слова, наиболее часто встречающегося в тексте); q_n – вероятность использования слова с рангом n; d и – константы. Известно, что существует взаимно однозначное преобразование между первым и вторым законами Ципфа: вывод первого из них влечет обоснование второго, и наоборот. Среди других исследований в этом направлении отметим математическую теорию эволюции Юла; цикл работ Чемпернауна [15] о закономерностях распределения доходов в экономике; подход, основанный на психофизиологических представлениях о деятельности головного мозга (А. Н. Лебедев) [16]; результат работы М. В. Арапова и др. в [16], полученный на основе точного учета целочисленности ранга слова.

Основной проблемой существующих аналитических моделей является наличие в них предположений, которые не всегда точно выполняются для исследуемых систем или процессов. Это несоответствие обычно указывается в качестве причины расхождений экспериментальных данных с результатами численного моделирования. Статистический анализ разнообразных данных (таких как частота слов в литературных произведениях на разных языках, индивидуальный и корпоративный доходы на примере различных государств, частота цитирования авторов и работ в научной литературе, численность персонала организаций и компаний, частота посещения различных Интернет-сайтов и Web-страниц и т. д. (см. обзор в [11, 12, 17])), выполненный многими авторами, вполне убедительно показывает, что закон Парето справедлив в основном в диапазоне больших k. Однако вместо поиска новых содержательных аналитических моделей исследования в данной области направлены, по сути, на подгонку результатов измерений (обычно путем предварительной фильтрации выборки) к одной из двух классических (Саймона или Мандельброта) моделей или же, наоборот, на подгонку этих моделей к данным путем ввода дополнительных параметров или компонент, которые не имеют какого-либо физического смысла [18]. Адекватная математическая модель должна, по всей видимости, учитывать, во-первых, дискретность случайной величины и, во-вторых, психофизиологический фактор формирования естественного языка.

В данной работе предлагается новая математическая модель механизма формирования распределения вероятностей частоты слова в текстах, таких как английский, русский, а также английский гипертекст (актуальная задача моделирования связи в Интернет-сети, которая не имеет решения в рамках классических подходов [19]). Получено распределение вероятностей частот, обобщающее закон Парето. Представлен алгоритм вычисления параметров модели. На примере разнообразных данных (смесь английских текстов, смесь гипертекстов, смесь текстов на русском языке) показано, что полученное распределение вероятностей обеспечивает существенно более высокую степень согласия с экспериментальными данными, чем распределение Парето.

1. Модель механизма формирования распределения вероятностей частоты слова. Для обозначения некоторого концепта (понятия) в естественных языках используется не единственное слово, а, как правило, множество слов (например, синонимы или слова, заимствованные из другого языка). Пусть p_k – вероятность использовать слово k раз для обозначения фиксированного концепта, т. е. p_k – условная вероятность события k при условии, что рассматривается только фиксированный концепт. Будем полагать, что среди всех распределений p_k с фиксированным математическим ожиданием $a \quad k p_k$ реализуется такое распределение, для которого энтропия $H \quad p_k \ln p_k$ достигает своего максимального значения. Поиск максимума функции H с условиями $p_k \quad 1$ и $k p_k \quad a$ сводится к задаче поиска безусловного максимума функции Лагранжа

$$p_k \ln p_k \quad 1 \quad p_k \quad a \quad k p_k \tag{1}$$

с множителями и , решение которой находится из системы уравнений: $/ p_k \quad 0, \quad / \quad 0, \quad / \quad 0$ [20]. Из первого и второго уравнений следует $p_k \quad P(\quad k/)$, где

$$P(\quad k/) \quad (e \quad 1)e^{\ k}, \quad k \quad 1,2,\ldots, \tag{2}$$

– условная вероятность события k при фиксированном . Таким образом, параметр является статистической характеристикой рассматриваемого концепта. Из (2) следует

$$a \quad k p_k \quad e \ /(e \quad 1) \quad 1/p_1.$$

Математическое ожидание a числа использований слова существенно больше единицы, поэтому параметры p_1 и существенно меньше единицы. При малом получаем $1/a$. Величина, обратная среднему числу использования слова, пропорциональна – среднему времени использования слова носителем языка для обозначения рассматриваемого концепта. Отметим, что экспоненциальное распределение частоты можно получить другим способом [21].

Параметр будем рассматривать как непрерывную случайную величину с плотностью ( ), следовательно, безусловное распределение вероятностей частоты использования слова имеет вид

$$P(\quad k) \quad ( )P(\quad k/)d \ , \quad k \quad 1,2,\ldots. \tag{3}$$

Для естественного языка измеряемая величина является смесью нескольких случайных величин, которые отражают влияние подсистем этого языка, например подсистемы служебных слов и подсистемы смысловых слов [22,

23]. При этом плотность распределения ( ) $\sum_{i\ 1}^{N} c_i\ _i(\ )$, где N – число компонент распределения (подсистем); c_i – вероятность обнаружить i-ю компоненту в наблюдаемой смеси; $_i(\ )$ – плотность вероятности случайной величины в i-й компоненте. Тогда

$$P(\quad k) \sum_{i\ 1}^{N} c_i P_i(k), \tag{4}$$

$$P_i(k) \quad _i(\quad P(\quad k/\ )d\ , \quad k\ \ 1,2,\ldots, \tag{5}$$
$$0$$

где $P_i(k)$ – распределение вероятности частоты использования слова в i-й компоненте.

Каждая из функций $_i(\ )$ – это, по сути, плотность вероятности величины – среднего времени использования слова для обозначения фиксированного концепта. Результаты многих физиологических экспериментов показывают, что линейно зависит от среднего времени реакции, которое является гамма-распределенной случайной величиной [24]. Поэтому полагаем $_i(\ )\ \ b_i{}^{\ i}\ {}^{\ i\ \ 1}\mathrm{e}^{\ b_i}\ /\ (\ _i)$, 0, где ( ) – гамма-функция, а $b_i\ \ 0$ и $_i\ \ 0$ – параметры распределения. Подстановка этого распределения и формулы (2) в (5) приводит к выражению

$$P_i(k)\ \ \frac{b_i{}^{\ i}}{(k\ \ 1\ \ b_i)^{\ i}}\ \ \frac{b_i{}^{\ i}}{(k\ \ b_i)^{\ i}}, \quad k\ \ 1,2,\ldots. \tag{6}$$

Таким образом, распределение вероятностей частоты использования слова определяется соотношениями (4), (6), которые получены на основе следующих положений предлагаемой математической модели механизма формирования частоты.

1. Для фиксированного концепта реализуется распределение вероятностей частоты использования слова с максимальным значением энтропии.

2. Среднее время использования слова для обозначения концепта зависит от среднего времени реакции и есть гамма-распределенная случайная величина.

3. Естественный язык как система передачи информации может состоять из нескольких статистически различимых подсистем.

2. Асимптотика распределения частоты. Если предположить, что соотношения (4), (6) хорошо согласуются с экспериментальными данными для всех значений $k\ \ 1,2,\ldots$, тогда в асимптотике при больших k эти соотношения должны переходить в распределение Парето, поскольку, как отмечалось, закон Парето хорошо согласуется с экспериментальными данными для больших k. Покажем, что действительно из соотношений (4), (6) при некоторых условиях следует распределение Парето. Пусть число компонент $N\ \ 1$, тогда $c_1\ \ 1$ и сумма (4) содержит единственное слагаемое $P(\quad k)\ \ P_1(k)$. При этом в формуле (6) у параметров b_i, $_i$ опустим индекс i, поскольку рассматривается единственное значение $i\ \ 1$. Пусть выполняется условие $k\ \ b\ \ 1$, тогда $z\ \ 1/(k\ \ b)\ \ 1$. Разлагая в ряд Тейлора по малому параметру

z функцию $(1 \quad z)$, имеем $(1 \quad z) \quad 1 \quad z \quad \ldots$. С учетом этого первое слагаемое (6) можно преобразовать следующим образом:

$$\frac{b}{(k \quad 1 \quad b)} \quad \frac{b}{(k \quad b)\ (1 \quad z)} \quad \frac{b}{(k \quad b)} \ 1 \quad \frac{}{k \quad b} \ . \tag{7}$$

Подставим (7) в (6), тогда

$$P_1(k) \quad \frac{b}{(k \quad b)^{\ 1}}, \quad k \quad b \quad 1. \tag{8}$$

Пусть выполняется дополнительное условие $k \quad b$, тогда из (8) следует распределение Парето $P_1(k) \quad b \ / k \quad ^1$. Отметим, что при $b \quad 1$ первому условию $k \quad b \quad 1$ удовлетворяют значения $k \quad 1$, а при $b \quad 1$ – значения $k \quad 1, 2, \ldots$. Второму условию $k \quad b$ при любом b удовлетворяют только значения $k \quad 1$. Поэтому второе условие более сильное, чем первое. Распределение Парето можно получить непосредственно из (6) при условии $k \quad b$. Для этого достаточно разложить в ряд Тейлора первое слагаемое (6) по малому параметру $(b \quad 1)/k$ и второе слагаемое по малому параметру b/k.

Соотношение (8) допускает следующую интерпретацию. Пусть рассматривается как непрерывная случайная величина, тогда аналогом формулы (2) будет экспоненциальная плотность распределения вероятностей $\mathrm{e}^{\ k}$, $k \quad 0$. При этом вычисление интеграла (5) приводит к соотношению (8). Таким образом, (8) можно рассматривать как плотность распределения вероятностей непрерывной случайной величины для $k \quad 0$. Такая плотность имеет максимальное значение $P_1(0) \quad /b$, является монотонной убывающей функцией и удовлетворяет условию нормировки $\int_0 P_1(k)\,dk \quad 1$. Пусть теперь имеются две плотности $P_1(k)$ и $P_2(k)$ вида (8) с параметрами $_1, b_1$ и $_2, b_2$ соответственно. Тогда условие $_1/b_1 \quad _2/b_2$ эквивалентно $P_1(0) \quad P_2(0)$, откуда следует, что при больших k должно выполняться обратное соотношение $P_1(k) \quad P_2(k)$, поскольку в противном случае нарушается условие нормировки. Это свойство функции (8) полезно для последующей интерпретации результатов анализа экспериментальных данных.

3. Оценки максимального правдоподобия. Проверка соотношений (4), (6) на соответствие экспериментальным данным заключается в оценивании параметров $N, b_1, \ldots, b_N, \ _1, \ldots, \ _N$ и последующем использовании критерия согласия теоретического распределения $P(\quad k)$ (с параметрами, равными их оценкам) и эмпирического распределения $W(k)$ частоты использования слова. Однако первая задача оценивания параметров распределения (4) в такой общей постановке очень сложна и вряд ли может быть решена известными методами теории оценивания [25]. Так, метод моментов здесь вообще неприменим, поскольку для распределения (6) существует небольшое число моментов, недостаточное для построения системы уравнений относительно оценок. Использование критерия ^{2} осложняется проблемой минимизации нелинейной функции $2N \quad 1$ переменных. Решение системы уравнений максимального правдоподобия становится проблематичным даже при $N \quad 1$. Эти

трудности приводят к необходимости поиска приближенных решений задачи оценивания.

Рассмотрим простой вариант приближенного решения в предположении, что распределение вероятностей частоты слова для $k \quad 1,2,\ldots$ задается соотношением (8). Имеются следующие доводы в пользу такого варианта.

1. Если верны соотношения (4), (6), тогда соотношение (8) верно для более широкого диапазона значений k, чем распределение Парето, так как последнее следует из (6) при более сильных ограничениях, чем (8). Поэтому можно ожидать, что соотношение (8) точнее описывает распределение частоты слова, чем распределение Парето.

2. Соотношение (8) имеет простые аналитические свойства, что позволяет получить практический алгоритм вычисления оценок максимального правдоподобия.

Пусть $x_1, \ldots, x_n$ – выборка объема n, полученная при выполнении n измерений случайной величины , тогда функция правдоподобия для распределения (8) имеет вид

$$\prod_{i \quad 1}^{n} P_1(x_i) \quad \prod_{i \quad 1}^{n} \frac{b}{(x_i \quad b)^{\quad 1}}. \tag{9}$$

Введем обозначение

$$L \quad \ln \prod_{i \quad 1}^{n} P_1(x_i) \quad \sum_{i \quad 1}^{n} [\ln \qquad \ln b \quad (\quad 1) \ln (x_i \quad b)] \tag{10}$$

и составим систему уравнений максимального правдоподобия: $L/ \quad 0$, $L/ \; b \quad 0$, решение которой будет определять оценки параметров , b. Пусть

$$t_1 \quad \frac{1}{n} \sum_{i \quad 1}^{n} \frac{b}{x_i \quad b}; \qquad t_2 \quad \frac{1}{n} \sum_{i \quad 1}^{n} \ln \frac{b}{x_i \quad b}, \tag{11}$$

тогда система уравнений максимального правдоподобия приводится к виду

$$\frac{1}{-} \qquad t_2; \qquad \frac{t_1}{1 \quad t_1}. \tag{12}$$

Исключив параметр из соотношений (12), получим

$$1 \quad t_1 \qquad t_1 t_2 \tag{13}$$

– уравнение относительно оценки параметра b, которое несложно решить численным методом. Подстановка этого решения в одно из соотношений (12) позволяет вычислить оценку параметра . Таким образом, выражения (12), (13) определяют решение системы уравнений максимального правдоподобия.

Для проверки соотношения (8) на соответствие экспериментальным данным была сформирована последовательность частоты использования слова в статьях политической тематики на английском языке, выбранных на сайте издательства “Associated Press”. При подсчете частоты слова игнорировались различия между заглавным и строчным регистрами в написании букв.

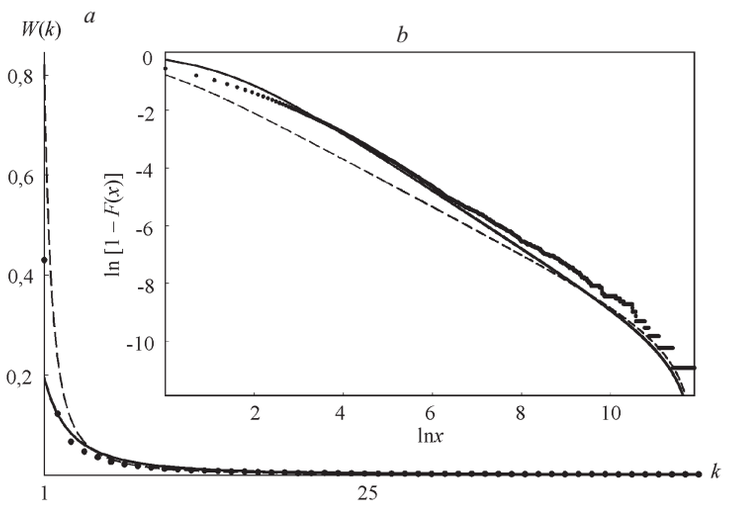

Рис. 1

Во всех экспериментах какой-либо иной предварительной обработки текстов, в том числе для учета морфологической изменчивости слов, не проводилось. Общий объем обработанных текстов составил 7,3 Мбайт при количестве слов 10^6. На рис. 1, а точками представлено эмпирическое распределение $W(k)$ частоты использования слова, пунктиром – распределение Парето и сплошной линией – распределение вида (8). Оценки максимального правдоподобия параметров распределения Парето были получены с использованием программы BestFit 4.5 “Palisade Corporation” для научных и инженерных статистических расчетов. На рис. 1, b в двойном логарифмическом масштабе представлены соответствующие интегральные функции распределения вероятностей для целого аргумента, например эмпирическая функция распределения $F(k) \quad \sum_{i \quad 1}^{k} W(i)$, $k \quad 1,2,\ldots$. Необходимо отметить, что наблюдаемое расхождение классической модели с эмпирическими данными хорошо известно в литературе [11] и не зависит от параметров текста, таких как объем, язык, однородность (в смысле, что это текст одного автора или подборка текстов разных авторов) и т. д.

4. Оценки параметров второй компоненты. Результаты анализа экспериментальных данных показывают, что соотношение (8) точнее описывает распределение вероятностей частоты слова, чем закон Парето, однако также не обеспечивает полного соответствия с эмпирическим распределением, особенно в области малых k. Видимо, это обусловлено тем, что естественный язык представляет собой смесь, как минимум, двух статистически различимых подсистем: служебных и смысловых слов [22]. Поэтому можно ожидать существенного повышения степени согласия эмпирического $W(k)$ и теоретического $P(\quad k)$ распределений, если использовать для последнего выражение (4) при $N \quad 2$. Статистическая различимость подсистем при этом означает существенное различие распределений $P_1(k)$ и $P_2(k)$ в формуле (4), что приводит к условию $P_1(k) \quad P_2(k)$ в области малых k и $P_1(k) \quad P_2(k)$ для больших k. Таким образом, при больших k из (4) следует $P(\quad k) \quad c_2 P_2(k)$.

Это позволяет сформулировать критерий выбора оценок параметров c_2, $_2$, b_2 в виде условия

$$\sum_{k \quad n_1}^{n_2} [\ln W(k) \quad \ln c_2 P_2(k)]^2 \qquad \min_{c_2, \ _2, b_2}. \tag{14}$$

Здесь число n_1 выбирается столь большим, чтобы при $k \quad n_1$ выполнялось неравенство $P_1(k) \quad P_2(k)$ и в соотношении (4) первым слагаемым можно было пренебречь. Второе число n_2 зависит от объема выборки и выбирается не слишком большим, так чтобы при $k \quad n_2$ эмпирическое распределение $W(k)$ существенно отличалось от нуля. При этом для $P_2(k)$ справедлива асимптотика (8). Подставим (8) в (14) и для упрощения обозначений в последующих преобразованиях опустим индексы у параметров c_2, $_2$, b_2, тогда

$$\sum_{k \quad n_1}^{n_2} [\ln W(k) \quad \ln c \quad \ln \qquad \ln b \quad (\quad 1) \ln(k \quad b)]^2. \tag{15}$$

Таким образом, задача оценивания параметров второй компоненты сводится к поиску минимума функции (15) трех аргументов c, $,b$. Значения аргументов в точке минимума принимаются за оценки параметров второй компоненты. Необходимые условия минимума функции нескольких аргументов в данном случае сводятся к системе трех уравнений [20]:

$$\frac{}{c} \quad 0; \qquad \frac{}{} \quad 0; \qquad \frac{}{b} \quad 0. \tag{16}$$

Пусть $n \quad n_2 \quad n_1 \quad 1$ – число слагаемых в сумме (15) и

$$z \quad \ln c \quad \ln \qquad \ln b. \tag{17}$$

Тогда первое уравнение системы (16) принимает вид

$$\frac{1}{n} \sum_{k \quad n_1}^{n_2} \ln W(k) \quad z \quad (\quad 1) \frac{1}{n} \sum_{k \quad n_1}^{n_2} \ln(k \quad b) \quad 0. \tag{18}$$

С учетом этого равенства второе уравнение системы (16) преобразуется к выражению

$$\frac{1}{n} \sum_{k \quad n_1}^{n_2} \ln W(k) \ln(k \quad b) \quad z \frac{1}{n} \sum_{k \quad n_1}^{n_2} \ln(k \quad b) \quad (\quad 1) \frac{1}{n} \sum_{k \quad n_1}^{n_2} \ln^2(k \quad b) \quad 0. \tag{19}$$

Третье уравнение (16) также с учетом равенства (18) сводится к выражению

$$\frac{1}{n} \sum_{k \quad n_1}^{n_2} \frac{\ln W(k)}{k \quad b} \quad z \frac{1}{n} \sum_{k \quad n_1}^{n_2} \frac{1}{k \quad b} \quad (\quad 1) \frac{1}{n} \sum_{k \quad n_1}^{n_2} \frac{\ln(k \quad b)}{k \quad b} \quad 0. \tag{20}$$

Соотношения (18)–(20) образуют систему трех уравнений относительно неизвестных z, , b, решение которой позволяет найти параметр с из равенства (17).

Введем обозначения:

$$s_1 \quad \frac{1}{n} \sum_{k \quad n_1}^{n_2} \frac{1}{k \quad b}; \qquad s_2 \quad \frac{1}{n} \sum_{k \quad n_1}^{n_2} \frac{\ln(k \quad b)}{k \quad b}; \qquad s_3 \quad \frac{1}{n} \sum_{k \quad n_1}^{n_2} \frac{\ln W(k)}{k \quad b}, \tag{21}$$

$$z_1 \quad \frac{1}{n} \sum_{k \quad n_1}^{n_2} \ln(k \quad b); \qquad z_2 \quad \frac{1}{n} \sum_{k \quad n_1}^{n_2} \ln^2(k \quad b), \tag{22}$$

$$z_3 \quad \frac{1}{n} \sum_{k \quad n_1}^{n_2} \ln W(k) \ln(k \quad b); \qquad z_4 \quad \frac{1}{n} \sum_{k \quad n_1}^{n_2} \ln W(k). \tag{23}$$

Выразим z из уравнения (18) и подставим его в (19) и (20), тогда в принятых обозначениях система трех уравнений имеет следующий вид:

$$z \quad z_4 \quad (\quad 1) z_1, \tag{24}$$

$$z_3 \quad z_1 z_4 \quad (\quad 1)(z_2 \quad z_1^2) \quad 0, \tag{25}$$

$$s_3 \quad s_1 z_4 \quad (\quad 1)(s_2 \quad s_1 z_1) \quad 0. \tag{26}$$

Из (25) находим

$$1 \quad \frac{z_3 \quad z_1 z_4}{z_1^2 \quad z_2}. \tag{27}$$

В соответствии с формулами (21)–(23) величины $s_1, s_2, s_3, z_1, z_2, z_3$ являются функциями аргумента b и z_4 – число. Поэтому соотношение (27) определяет параметр как функцию аргумента b. Подставим (27) в (26) и получим уравнение

$$s_3 \quad s_1 z_4 \quad (s_2 \quad s_1 z_1)(z_3 \quad z_1 z_4) / (z_1^2 \quad z_2) \quad 0 \tag{28}$$

относительно неизвестного b. Его решение является оценкой параметра b_2. При $b \quad b_2$ соотношение (27) определяет оценку параметра $_2$, подстановка $_2$, $b \quad b_2$ в (24) – величину z, и затем из соотношения (17) вычисляется оценка c_2.

5. Оценки параметров первой компоненты. Подставим полученные параметры $_2$, b_2 в формулу (6) и вычислим $P_2(k)$ для $k \quad 1,2,\ldots$. Затем найдем

$$W_1(k) \quad \frac{W(k) \quad c_2 P_2(k)}{1 \quad c_2}, \quad k \quad 1,2,\ldots, \tag{29}$$

– эмпирическое распределение частоты слова первой компоненты. Здесь $1 \quad c_2 \quad c_1$ – оценка вероятности того, что слово принадлежит первой компоненте. Приравнивая теоретическое (6) и эмпирическое распределения

$P_1(k)$ $W_1(k)$ для k 1,2, получаем систему двух уравнений относительно неизвестных , b. Из первого уравнения (k 1) следует

$$\frac{\ln[1 \quad W_1(1)]}{\ln b \quad \ln(b \quad 1)}. \tag{30}$$

Подставим (30) во второе уравнение (k 2) и получим

$$\frac{b}{b \quad 1}^{\frac{\ln[1 \quad W_1(1)]}{\ln b \quad \ln(b \quad 1)}} \quad \frac{b}{b \quad 2}^{\frac{\ln[1 \quad W_1(1)]}{\ln b \quad \ln(b \quad 1)}} \quad W_1(2) \tag{31}$$

– уравнение относительно b. Его решение принимаем за оценку параметра b_1. Подстановка b b_1 в соотношение (30) определяет оценку параметра $_1$.

Представленный алгоритм использовался для вычисления параметров теоретического распределения вида (4), (6) по той же выборке частот, которая рассматривалась в разд. 3. На рис. 2, а показаны эмпирическое распределение $W(k)$ (точки) и теоретическое распределение $P($ $k)$ (сплошная линия), а на рис. 2, b – эмпирическая функция распределения $F(k)$ (точки) и теоретическая функция распределения для целого аргумента (сплошная линия).

Такие же результаты получены для английского гипертекста (выборка Web-страниц сайта www.cbsnews.com, сформированная случайным образом; объем 11,4 Мбайт, число слов 1,8 10^6) и текста на русском языке (подборка коротких литературных произведений разных авторов, произвольным образом выбранных на сайте www.lib.ru; объем 3,1 Мбайт, число слов 0,45 10^6). Основные параметры c_1, c_2, $_1/b_1$, $_2/b_2$ распределения частоты для английского текста равны 0,525, 0,475, 0,805, 0,136; для гипертекста – 0,549, 0,451, 0,572, 0,123; для русского текста – 0,673, 0,327, 0,978, 0,363 соответственно. Отметим, что для английского текста и гипертекста расчетные значения вероятностей c_1 и c_2 хорошо согласуются с известной

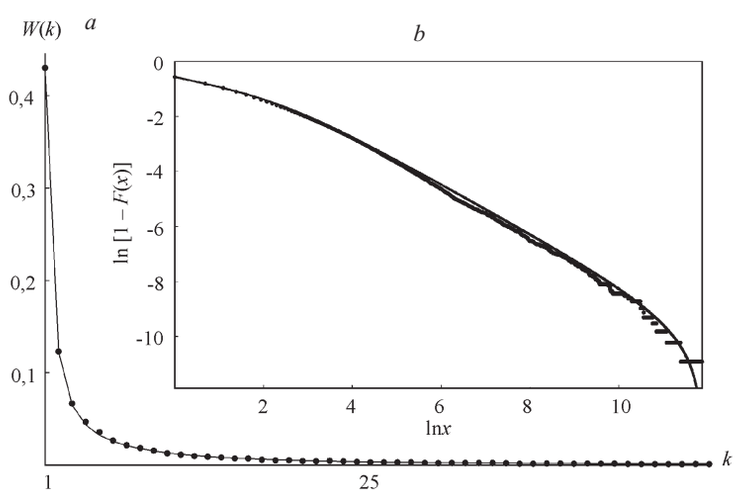

Рис. 2

статистикой смысловых и служебных слов, на которые в современном английском языке приходится соответственно 50–55 % и 45–50 % объема текста [22, 11]. Этот факт представляется весомым аргументом в пользу предлагаемой модели.

Отметим, что критерий оптимальности (14) – это частный случай метода наименьших квадратов [26, 27], где функция $\ln c_2 P_2(k)$ дифференцируема по всем параметрам. Для большого объема выборки случайные остатки $\ln W(k) \quad \ln c_2 P_2(k)$ имеют нулевые средние и одинаковые дисперсии. Поэтому условие (14) определяет состоятельные оценки. Оценки параметров первой компоненты (29)–(31) также состоятельны, поскольку вычисляются по методу моментов: приравниванием эмпирического и теоретического распределений вероятностей для $k \quad 1,2$.

Заключение. В работе получено распределение вероятностей частоты использования слова в текстах, обобщающее закон Парето, а также разработан алгоритм вычисления его параметров. На примере разнообразных данных показано, что полученное распределение вероятностей обеспечивает существенно более высокую степень согласия с экспериментальными данными, чем распределение Парето. Этот результат представляется принципиально важным с точки зрения создания разнообразных сетевых мультимедийных приложений [19], а также систем поиска информации в Интернет-сети [28], так как полученная аналитическая модель позволяет с хорошей точностью вычислить важнейшие характеристики процесса передачи информации. Можно ожидать, что полученное распределение будет полезно для анализа работы телекоммуникационных сетей [9, 29, 30] и различных социальных, в том числе экономических, систем [10, 31].

СПИСОК ЛИТЕРАТУРЫ

1. Van Rijsbergen C. J. Information Retrieval. London: Butterworths, 1979.
2. Losee R. M. Term dependence: A basis for Luhn and Zipf models // Journ. Amer. Soc. Inform. Sci. and Technol. 2001. 52, N 12. P. 1019.
3. Brin S., Page L. The anatomy of a large-scale hypertextual Web search engine // Comput. Networks and ISDN Systems. 1998. 30, N 1–7. P. 107.
4. Peters C., Koster C. Uncertainty and term selection in text categorization // Intern. Journ. Uncertainty, Fuzziness and Knowledge-Based Systems. 2003. 11, N 1. P. 115.
5. Amati G., Van Rijsbergen C. J. Probabilistic models of information retrieval based on measuring the divergence from randomness // ACM Trans. on Inform. Systems (TOIS). 2002. 20, N 4. P. 357.
6. Amati G., Van Rijsbergen C. J. Term frequency normalization via Pareto distributions // Springer Lecture Notes in Comput. Sci. 2002. 2291. P. 183.
7. Pareto V. Cours d'Economie Politique. Lausanne – Paris: Rouge, 1897.
8. Zipf G. K. Human Behavior and the Principle of Least Effort. Cambridge, Mass.: Addison-Wesley, 1949.
9. Barabasi A., Albert R. Emergence of scaling in random networks // Sci. Magazine. 1999. 286. P. 509.
10. Axtell R. L. Zipf distribution of U.S. firm sizes // Sci. Magazine. 2001. 293. P. 1818.
11. http://www.ram-verlag.de (Glottometrics, Special issue to honor G. K. Zipf. 2002. N 4.)

12. Mitzenmacher M. A brief history of generative models for power law and lognormal distributions // Internet Mathem. 2003. 1, N 2. P. 226.
13. Simon H. A. On a class of skew distribution functions // Biometrika. 1955. N 42. P. 425.
14. Mandelbrot B. B. An informational theory of the statistical structure of languages // Commun. Theory. Betterworth. 1953. P. 486.
15. Champernowne D. The Distribution of Income. Cambridge: Cambridge University Press, 1973.
16. Тулдава Ю. А. Проблемы и методы квантитативно-системного исследования лексики. Таллин: Валгус, 1987.
17. Shide N., Batty M. Power Law Distributions in Real and Virtual Worlds // http://www.isoc.org/inet2000 (Inet 2000 Proc., Internet Soc. 2000.)
18.
19.
20. Бронштейн И. Н., Семендяев К. А. Справочник по математике. М.: Наука, 1986.
21. Kryssanov V. V., Kakusho K., Kuleshov E. L., Minoh M. Modeling hypermedia-based communication // Intern. Journ. Inform. Sci. 2004; http://authors.elsevier.com/sd/article/S0020025504002385
22. Balasubrahmanyan V. K., Naranan S. Quantitative linguistics and complex system studies // Journ. Quant. Linguistics. 1996. N 3. P. 177.
23. Naranan S., Balasubrahmanyan V. K. Information theoretic models in statistical linguistics. Pt. I, II // Current Sci. 1992. N 63. P. 261, 297.
24. Luce R. D. Response Times. Their Role in Inferring Elementary Mental Organization. New York: Oxford University Press, 1986.
25. Королюк В. С., Портенко Н. И., Скороход А. В., Турбин А. Ф. Справочник по теории вероятностей и математической статистике. М.: Наука, 1985.
26. Рао С. Р. Линейные статистические методы и их применения. М.: Наука, 1968.
27. Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика. Основы моделирования и первичная обработка данных. М.: Финансы и статистика, 1983.
28. Bates M. Indexing and access for digital libraries and the Internet: Human, database, and domain factors // Journ. Amer. Soc. Inform. Sci. 1998. 49, N 13. P. 1185.
29. Huberman B. A., Pirolli P. L. T., Pitkow J. E., Lukose R. M. Strong regularities in World Wide Web surfing // Sci. Magazine. 1998. N 280. P. 95.
30. Adamic L. A., Huberman B. A. Power-law distribution of the World Wide Web // Sci. Magazine. 2000. 287. P. 2115a.
31. Stanley M. H. R., Amaral L. A. N., Buldyrev S. V. et al. Scaling behavior in the growth of companies // Nature. 1996. 379. P. 804.

Дальневосточный государственный университет,
Рицумеикан университет, Киото университет,
E-mail: kuleshov@lemoi.phys.dvgu.ru

Поступила в редакцию
8 июня 2004 г.