УДК 519.6 **DOI** 10.15622/SP.50.8

Ю.А. Котов

АППРОКСИМАЦИЯ РАСПРЕДЕЛЕНИЙ ЧАСТОТ БУКВЕННЫХ БИГРАММ ТЕКСТА ДЛЯ ИДЕНТИФИКАЦИИ БУКВ

Котов Ю.А. **Аппроксимация распределений частот буквенных биграмм текста для идентификации букв.**

Аннотация. В статье рассмотрены особенности применения методов частотного упорядочивания и аппроксимации для решения задачи идентификации знаков текста. Определены условия реализации метода Якобсена для получения наименьшей погрешности идентификации. Предложен метод аппроксимации одномерных и двумерных распределений частот знаковых биграмм текста и буквенных биграмм эталона языка текста. Приведены экспериментальные данные о погрешностях метода Якобсена и предложенного метода аппроксимации для русскоязычных текстов.

Погрешность предложенного метода меньше, чем у метода Якобсена. Метод может быть использован для идентификации знаков текста любого языка, для которого существует эталонное распределение частот буквенных биграмм.

Ключевые слова: аппроксимация, идентификация, буква, биграмма, простая замена, шифр.

1. Введение. Алфавит любого языка представляет собой множество упорядоченных кодовых знаков, обозначающих буквы этого языка. Буквы языка однозначно связаны с их порядковым номером в исходном алфавите, но могут быть представлены разными — в том числе неизвестными — знаками, и в то же время одинаковые знаки в разных текстах могут обозначать одну и ту же, но, возможно, неизвестную букву. Будем считать, что в каждом конкретном тексте для обозначения буквы языка используется только один знак. Тогда идентифицировать знак в произвольном тексте на некотором языке означает приписать ему такие числовые характеристики, получаемые из данного текста, которые позволяют определить номер знака в исходном алфавите этого языка и соответственно букву, которую данный знак представляет [1].

Сложность решения данной задачи заключается в том, что каждый текст имеет собственное упорядочение множества используемых в нем знаков. Математически это означает, что каждый текст определен в своей собственной системе координат. Для формального анализа дискретных множеств, которые представляют тексты и их элементы, необходимо преобразование индивидуальных систем координат каждого текста к общей системе. В большинстве случаев такое преобразование осуществляется отождествлением множеств используемых в тексте знаков, «букв», с некоторым известным алфавитом. Если же отождествление невозможно, тексты признаются несравнимыми, хотя, возможно, отличаются только множествами используемых знаков.

Традиционно задача идентификации знаков таких текстов трактуется как криптографическая, связанная с шифром простой замены [2]. Для решения данной задачи могут применяться методы перебора, частотного упорядочивания [3–7], генетические алгоритмы [9-11], марковские цепи [12–13], методы, основанные на вычислении энтропии [2, 14–15] и статистической проверке гипотез [16], а также другие методы и их комбинации, например, [1, 10, 13, 17]. В основе большинства этих методов лежат методы частотного упорядочивания, представляющие частный случай методов аппроксимации.

2. Методы упорядочивания и аппроксимации в задаче идентификации знаков. Идея использования метода частотного упорядочения для решения задачи идентификации знаков основана на следующих математических принципах [3-6]. Пусть есть множество пар знаков эталона и частоты появления этих знаков $\mathbf{E}=\{<e_i, h_i>\}$ и аналогичное множество $\mathbf{T}=\{<z_j, k_j>\}$ для знаков анализируемого текста T. Соответствие между знаками **E** и **T** заранее неизвестно. Так как знаки имеют нечисловую природу, построить функции для множеств **E** и **T** можно только на основании значений частот знаков, то есть значений из области изменения функций.

Для решения этой задачи предположим, что знаки множеств **E** и **T** подчиняются одинаковым функциональным закономерностям и в любой системе координат имеют одинаковый знак приращения функции в каждой точке области определения, а в области изменения функции нет совпадающих значений. Тогда в системе координат, в которой значения частот знаков эталона **E** упорядочены (по убыванию или возрастанию), частоты знаков **T** будут упорядочены таким же образом. Построим в этой системе функции $y=f_E(x)$ и $y=f_T(x)$, применив одинаковый способ упорядочивания к элементам множеств **E** и **T** и приписав в результате знакам из **E** и **T** значения из множества $\mathbf{N}=\{0,1,2,\ldots,n\text{-}1\}$, $x\in\mathbf{N}$, где n — мощность множеств **E** и **T**.

Суммарная разность (1):

$$W(\mathbf{T})=\sum_{i=0}^{n-1}\left|f_E(i)-f_T(i)\right|, \qquad (1)$$

значений $f_E(x)$ и $f_T(x)$ будет минимальной по построению. Предполагается, что $W(T)$ стремится к нулю при росте объема анализируемого текста, то есть в пределе функции $f_T(x)$ и $f_E(x)$ совпадают. По результатам одинакового упорядочения знаки анализируемого текста z_i отождествляются с соответствующими знаками эталона e_i , $z_i \equiv e_i$.

Погрешность такого решения определяется количеством знаков из **T**, правильно отождествленных со знаками эталона **E**, и зависит от соответствия эталона и анализируемого текста исходным посылкам метода, способа формирования эталона и его шкалы измерения, а также от объема анализируемого текста.

Значения эталона зависят не только от общего объема текстов, из которого он сформирован, но также от количества и разнообразия этих текстов, то есть от того, насколько в эталоне учитывается влияние стилистики тематической области текстов и их авторов. Очевидно следующее замечание.

Замечание 1. Наименьшую погрешность методов идентификации знаков текста, использующих сравнение с эталоном, можно ожидать в случаях, когда анализируемый текст является фрагментом текста, из которого сформирован эталон.

Таким образом, наиболее объективную информацию о методах идентификации знаков текста, использующих эталон, можно получить, когда анализируемый текст и эталон сформированы независимо друг от друга и представляют разные тексты. Замечание 1 необходимо, потому что в ряде работ, посвященных подобным методам, в частности [9-13], результаты работы методов представляются на фрагментах текста и эталонах, полученных из одного текста.

Группа методов частотного упорядочивания тесно связана с методами аппроксимации, а точнее — *поиска наилучшей* аппроксимации, применяемыми для решения задачи идентификации знаков текста. Эта связь определяется функцией $W(\mathbf{T})$. В терминах методов аппроксимации данная функция может рассматриваться как погрешность аппроксимации функции $f_E(x)$ функцией $f_T(x)$. Так как функция $W(\mathbf{T})$ — экстремальная функция, то ее можно взять в качестве целевой функции и осуществить поиск минимального значения $W(\mathbf{T})$, то есть решить оптимизационную задачу.

Тогда, казалось бы, можно отказаться от упорядочивания вообще, и сформулировать задачу идентификации знаков текста как задачу поиска наилучшей в смысле минимума функции $W(\mathbf{T})$ аппроксимации произвольной эталонной функции $f(E)$.

Однако справедливо не только то, что для одномерного случая подобная сложность решения избыточна, но и следующее замечание.

Замечание 2. Минимальное значение $W(\mathbf{T})$ не является единственным для различных упорядочений **T**.

Это легко показать для монотонных функций $f_E(x)$ и $f_T(x)$**.**

Отсюда следует, что в общем случае погрешность решения задачи идентификации знаков текста на основе метода аппроксимации будет больше при использовании неупорядоченных множеств **E** и **T.**

Справедливо также и то, что любой метод частотного упорядочивания является частным случаем метода аппроксимации.

3. Метод Якобсена. Непосредственно для решения задачи идентификации знаков текста (в формулировке раскрытия шифра простой замены) известен метод Якобсена [7]. Метод основан на минимизации значения целевой функции:

$$W(\mathbf{T}) = \sum_{ij} \left| t_{ij} - b_{ij} \right|, \qquad (2)$$

где t_{ij}, b_{ij} — значения частот знаковых (буквенных) биграмм анализируемого и эталонного текстов.

В одномерном случае по формуле (1) $W(\mathbf{T})$ может быть интерпретирована как площадь, ограниченная кривыми, аппроксимирующими значения частот множеств **E** и **T,** а поиск наилучшей аппроксимации — как поиск наилучшего совпадения форм этих кривых. В отличие от этого в двумерном случае, описываемом формулой (2), ведется поиск наилучшей аппроксимации поверхности $z_B=f_B(x,y)$, аппроксимирующей множество частот биграмм эталона **B,** поверхностью $z_T=f_T(x,y)$, аппроксимирующей множество частот биграмм **T** анализируемого текста. В этом случае $W(\mathbf{T})$ можно интерпретировать как объем, заключенный между соответствующими поверхностями.

Множества **B** и **T** образованы тройками $\mathbf{B}=\{<<e_i,e_j>,b_{ij}>\}$, $\mathbf{T}=\{<<w_i,w_j>,t_{ij}>\}$, где e_i , w_i — знаки, b_{ij} , t_{ij} — частоты биграмм эталона и анализируемого текста соответственно. Они могут быть представлены симметрично упорядоченными по знакам таблицами (матрицами), по строке которых идентифицируется первый знак биграммы, по столбцу — второй. Значениями таблицы являются значения соответствующих частот b_{ij} или t_{ij} .

Поиск минимального значения $W(\mathbf{T})$ в методе Якобсена осуществляется следующим образом [7]. По таблице биграмм **T** последовательно просматриваются пары знаков — сначала соседние, затем — отстоящие на один знак и так до тех пор, пока не будет рассмотрена пара, состоящая из первого и последнего знака таблицы. Назовем просмотр одинаково отстоящих пар *уровнем* просмотра, индекс пары на уровне — ее *местом*. Для каждого уровня просмотр ведется с начала таблицы.

При каждом просмотре пары проверяется необходимость ее перестановки в смысле минимизации значения $W(\boldsymbol{T})$. Для этого по формуле (2) вычисляется $W_0(kl)$ для исходной пары (k — уровень пары,

l — место), и $W_1(kl)$ — для *условно* переставленной. Если выполняется условие (3):

$$W_0(kl) > W_1(kl), \tag{3}$$

то пара переставляется, и просмотр начинается *с начала с нулевого уровня.*

Так продолжается до тех пор, пока при очередном просмотре пар с нулевого уровня не будет найдено ни одной перестановки, включая последнюю пару.

Метод Якобсена представляет собой обыкновенную сортировку с переменным шагом, адаптированную к решению задачи идентификации знаков текста на основе двумерного распределения частот знаковых биграмм и функции (3). Этот метод является методом двумерного упорядочивания, аналогичным одномерному методу упорядочивания частот знаков. Основным его недостатком, как и других методов упорядочивания, является возможность остановки в локальном минимуме целевой функции [8].

Метод Якобсена является классическим и эффективным представителем группы методов аппроксимации, применяемых для решения задачи идентификации знаков текста. Представляется очевидным его применение в качестве эталонного метода для сравнения с другими методами этой группы.

В работе [7] проверка метода была проведена для англоязычного текста «Alice in Wonderland» с использованием эталона биграмм, сформированного по тексту «Moby Dick». В данной работе была проведена экспериментальная проверка метода Якобсена на представительных выборках текстовых фрагментов русскоязычных текстов (раздел 5) с использованием независимого эталона биграмм из [16]. По результатам эксперимента можно сделать следующие замечания.

Метод Якобсена имеет низкую погрешность идентификации знаков, но значения ее существенно выше, чем это заявлено автором — например, не 2% при объеме текста в 1000 знаков [7], а 22% — 38% (таблицы1 и 2, рисунки1 и 2).

При этом наименьшую погрешность удается достичь только при замене условия (3) условием (4):

$$W_0(kl) - W_1(kl) > \varepsilon, \quad \varepsilon = 10^{-8}, \tag{4}$$

то есть только при учете погрешности вычислений. При других значениях ε, в том числе равном нулю, как в условии (3), погрешность метода значительно возрастает, иногда в несколько раз.

И наконец погрешность метода также возрастает, если матрицы **B** и **T** не упорядочены по частоте появления знаков эталона и анализируемого текста соответственно (объяснение см. в замечании 2 раздела 2) (необходимость начальной упорядоченности указана в [7]).

Данные о погрешности метода Якобсена, реализованного с учетом указанных замечаний, применительно к русскоязычным текстам приведены в таблицах 1 и 2 и на рисунках 1 и 2 раздела 5.

4. Описание метода аппроксимации. При использовании для идентификации знаков текста двумерных распределений знаковых биграмм, как в методе Якобсена [7], следует заметить, что из общего распределения биграмм для каждого знака эталона и анализируемого текста можно выделить два индивидуальных *условных* распределения биграмм, начинающихся и заканчивающихся данным знаком. Это одномерные распределения, получаемые в своеобразных «сечениях» общего двумерного распределения.

Тогда для идентификации знаков можно было бы применить метод Пирсона проверки однородности индивидуальных распределений [18], если бы не ограничения, связанные с тем, что такие распределения определены в различных системах координат для эталона и анализируемого текста.

Возьмем из метода Пирсона способ аппроксимации распределений, но модифицируем сам метод следующим образом:

1) оценку аппроксимации по распределению хи-квадрат Пирсона заменим прямой оценкой погрешности аппроксимации, учитывающей пропущенные значения и влияние «хвостов» сравниваемых распределений;

2) аппроксимацию распределений для каждого знака анализируемого текста будем проводить на всем множестве индивидуальных распределений частот биграмм знаков («букв») эталона;

3) идентификацию знаков анализируемого текста и образца будем осуществлять на основе минимальной оценки погрешности аппроксимации индивидуальных распределений;

4) введем специальную процедуру разрешения коллизий идентификации;

5) общую погрешность идентификации знаков анализируемого текста и образца будем оценивать на основе формулы (2).

Эти модификации приводят к следующему методу аппроксимации.

Пусть **AB** и **AT** — множества знаков («букв»), встречающихся в эталонном и анализируемом текстах соответственно, упорядоченные по частоте появления этих знаков. Назовем номер знака в этой упорядоченности его идентификатором. $VB=|\mathbf{AB}|$ и $VT=|\mathbf{AT}|$ — количества знаков

для множеств **AB** и **AT** соответственно, $VT \le VB$. Множества **AB** и **AT** одновременно либо содержат, либо не содержат знак «пробел».

Пусть **B** — таблица значений частот буквенных биграмм эталонного текста (образец), симметрично упорядоченная в соответствии с **AB, T** — таблица значений частот знаковых биграмм анализируемого текста (тест), симметрично упорядоченная в соответствии с **AT.** Диагональные элементы в обеих таблицах обнулены. С учетом этих соглашений по упорядоченности, будем далее под **B** и **T** понимать соответствующие квадратные матрицы значений частот биграмм $b_{ij} \in \mathbf{B}$, $t_{ij} \in \mathbf{T}$.

1. Для каждого элемента $a_m^T \in \mathbf{AT}$ последовательно вычислим по формулам (5) два вектора $\mathbf{S}_m^1$ и $\mathbf{S}_m^2$ по строкам и столбцам матриц **B** и **T** соответственно:

$$s_{mj}^1 = \sum_{i=1}^{VB} \frac{(t_{mi} - b_{ji})^2}{t_{mi} + b_{ji}}, \qquad s_{mj}^2 = \sum_{i=1}^{VB} \frac{(t_{im} - b_{ij})^2}{t_{im} + b_{ij}} \tag{5}$$

При этом если знаменатель в формулах (5) равен нулю, то значение соответствующей дроби в сумме принимается равным нулю.

Из (5) следует, что каждая строка (столбец) теста **T** сравниваются со всеми строками (столбцами) эталона **B**.

2. Для каждого элемента $a_m^T \in \mathbf{AT}$ по формулам (6) вычислим нормированную обратную ошибку аппроксимации для векторов $\mathbf{S}_m^1$ и $\mathbf{S}_m^2$ — векторы $\mathbf{R}_m^1$ и $\mathbf{R}_m^2$ соответственно:

$$r_{mj}^1 = 1 - \frac{s_{mj}^1}{\sum_{i=1}^{VB} b_{ji} + \sum_{i=1}^{VT} t_{mi}}, \qquad r_{mj}^2 = 1 - \frac{s_{mj}^2}{\sum_{i=1}^{VB} b_{ij} + \sum_{i=1}^{VT} t_{im}} \tag{6}$$

и вычислим общую ошибку $\mathbf{R}_m = \mathbf{R}_m^1 + \mathbf{R}_m^2$.

Критерий (6) сформирован на основе простых геометрических представлений о близости форм подобных кривых и линейной интерполяции пропущенных значений, и на основе формул (1) и (5).

3. Упорядочим векторы $\mathbf{R}_m$ по убыванию значений r_{mk}, с сохранением значения первоначального упорядочивания k. Значение k является идентификатором знака («буквы») эталона $a_k^B \in \mathbf{AB}$, и таким образом по значению r_{mk} устанавливается упорядоченная пара $<m,k>$,

идентифицирующая знак теста $a_m^T \in \mathbf{AT}$ на множестве знаков эталона **AB.** В качестве такой пары выберем первую пару из упорядоченного вектора $\mathbf{R}_m$.

4. Так как каждый знак $a_m^T \in \mathbf{AT}$ идентифицируется независимо от других знаков **AT** на всем множестве знаков эталона **AB**, то при выполнении шага 3 могут возникать коллизии, когда для одного и того же знака эталона выбираются два и более знака теста.

Для разрешения коллизий осуществляется следующая процедура «выталкивания». Если пара <*m,k*> уже идентифицирована, но появляется «претендент» *j*, то выбирается пара <*m,k*> или <*j,k*>, у которой больше значение r_{mk} или r_{jk} и происходит новая идентификация. Выбывший знак переходит на место, определяемое в порядке убывания ошибки из собственного упорядоченного вектора $\mathbf{R}_m$ или $\mathbf{R}_j$ («выталкивается»). Образуется тройка <*m,i,k*> означающая, что знак $a_m^T \in \mathbf{AT}$ идентифицирован как знак $a_k^B \in \mathbf{AB}$ по значению $r_{mk}^i \in \mathbf{R}_m$, $i = \overline{1, VB}$. Процедура продолжается, пока все коллизии не будут разрешены.

Заметим, что процедура выталкивания не эквивалентна простому упорядочиванию максимальных значений.

5. Выбираем следующий элемент $a_{m+1}^T \in \mathbf{AT}$ и переходим к шагу 3 и так. до полной идентификации всех знаков анализируемого текста.

6. Шаги 1-5 дают *частное* решение $\mathbf{U}_n$ исходной задачи, $n \geq 1$, в соответствии с которым таблица **T** переупорядочивается и образуется таблица $\mathbf{T}_n$.

Погрешность полученного решения $W(\mathbf{T}_n)$ вычисляется по формуле (2) для матриц **B** и $\mathbf{T}_n$. Шаги 1-5 повторяются до тех пор, пока выполняется условие (7):

$$W(\mathbf{T}_{n-1}) - W(\mathbf{T}_n) > \varepsilon, \quad \varepsilon = -0.2, \quad \mathbf{T}_0 = \mathbf{T}, \quad n < 20, \qquad (7)$$

аналогичное условию (4). В отличие от условия (4), условием (7) допускается увеличение $W(\mathbf{T}_n)$ относительно $W(\mathbf{T}_{n-1})$ и совпадающие значения, поэтому вводится ограничение на количество итераций n . Это позволяет пропускать локальные минимумы целевой функции при поиске глобального экстремума и ограничивать поиск в случаях неудачи. Значения ε и ограничения для n получены в результате вычислительного эксперимента. Если при получении очередного решения $\mathbf{U}_n$ не

было осуществлено ни одной перестановки по сравнению с $\mathbf{U}_{n-1}$, то итерационный процесс (7) прекращается независимо от значения n.

В качестве окончательного результата работы метода принимается результат предпоследней итерации — решение $\mathbf{U}_{n-1}$.

Возможность с помощью итерационного процесса (7) управлять получаемой погрешностью наряду с использованием аппроксимации одномерных условных распределений биграмм для идентификации знаков текста является основной особенностью рассмотренного метода аппроксимации, отличающей его от метода Якобсена.

Так же как и метод Якобсена, рассмотренный метод аппроксимации позволяет идентифицировать знаки в текстах на любых языках, для которых существуют эталонные распределения частот буквенных биграмм, то есть является универсальным.

5. Результаты вычислительного эксперимента. Экспериментальная проверка погрешности метода Якобсена и предложенного метода аппроксимации была проведена на двух выборках фрагментов русскоязычных текстов.

Выборка 1 была сформирована из 100 научно-популярных и художественных текстов разных жанров и авторов; выборка 2 — из 100 текстов учебных пособий для вузов разных авторов из различных областей знаний: математика, химия, физика, машиностроение и т.д.

Из текстов 1-го и 2-го типов случайным образом были выделены последовательные фрагменты различной длины.

Выделение фрагментов для выборки 1 происходило после удаления из текста всех пробелов. Выделение происходило по принципу вложенности: сначала выделялись фрагменты большей длины, затем из них выделялись меньшие фрагменты последовательным удалением постоянного объема знаков. При этом начала фрагментов для одного текста совпадали. Это означает, что фрагменты разной длины для одного текста в выборке 1 включены друг в друга, а длина фрагментов совпадает со шкалой длин, представленной в эксперименте.

Фрагменты для выборки 2 выделялись со случайного знака текста с учетом пробелов, которые после этого удалялись. Таким образом, в отличие от выборки 1, выборка 2 сформирована из фрагментов случайной длины, начинающихся со случайного знака текста. Пересечения фрагментов для одного текста в выборке 2 возможны только случайно. Полученные для данных фрагментов значения ошибок отнесены к ближайшим значениям *сверху* шкалы длин фрагментов, одинаковой со шкалой выборки 1.

Выборки фрагментов для разных групп производились независимо друг от друга. Фрагменты обеих выборок распределены по 4

группам: группа 1 — фрагменты длиной от 400 до 1800 знаков, с шагом 200, группа 2 — фрагменты от 2000 до 10000 знаков, с шагом 2000, группа 3 — фрагменты от 30000 до 90000 знаков, с шагом 20000, группа 4 — фрагменты от 100000 до 350000 знаков, с шагом 50000.

Таблица 1. Результаты вычислительного эксперимента 1

N	*K*	*O*1*m*	*O*1*j*	*O*2*m*	*O*2*j*	*max O*2*m*	*max O*2*j*	*SD O*2*m*	*SD O*2*j*
Группа 1									
400	20	20	20	12.30	16.10	24	27	4.95	5.47
600	49	45	47	9.44	9.74	22	24	5.46	5.51
800	64	57	57	8.47	7.25	25	26	5.77	5.18
1000	76	64	62	6.30	6.97	23	24	4.46	6.34
1200	82	56	49	4.50	4.31	19	22	3.26	4.10
1400	86	53	50	3.83	4.30	17	20	3.23	3.34
1600	90	41	41	3.00	4.17	13	15	1.95	3.12
1800	93	35	26	2.86	3.88	9	15	1.68	2.86
Группа 2									
2000	93	32	26	2.59	3.12	6	12	1.17	2.06
4000	99	12	13	2.08	2.38	3	4	0.28	0.62
6000	100	7	9	2.43	3.00	3	5	0.49	1.15
8000	100	3	5	2.67	3.40	3	5	0.47	0.80
10000	100	3	6	3.33	2.83	5	3	1.25	0.37
Группа 3									
30000	50	2	1	2.00	3.00	2	3	0.00	0.00
50000	50	1	1	2.00	2.00	2	2	0.00	0.00
70000	50	1	1	2.00	2.00	2	2	0.00	0.00
90000	50	0	1	0.00	2.00	0	2	0.00	0.00
Группа 4									
100000	20	0	0	0.00	0.00	0	0	0.00	0.00
150000	20	0	0	0.00	0.00	0	0	0.00	0.00
200000	20	0	1	0.00	3.00	0	3	0.00	0.00
250000	20	0	0	0.00	0.00	0	0	0.00	0.00
300000	20	0	1	0.00	3.00	0	3	0.00	0.00
350000	20	0	0	0.00	0.00	0	0	0.00	0.00

Всего в эксперименте использовался 2631 фрагмент, из которых выборка 1 содержит 1372 фрагмента (группа 1 — 560, группа 2 — 492, группа 3 — 200, группа 4 — 120), а выборка 2 — 1259 фрагментов (группа 1 — 547, группа 2 — 481, группа 3 — 183, группа 4 — 48). При этом в каждом из фрагментов использовались все буквы русского алфавита (31, «Е» — «Е,Ё», «Ь» — «Ь,Ъ»).

В эксперименте определялось 2 вида ошибок: $O1$ — ошибка полной идентификации, как количество ошибочных фрагментов, содержащих хотя бы один неправильно идентифицированный знак; $O2$ — ошибка идентификации знаков, как среднее количество неправильно идентифицированных знаков для всех ошибочных фрагментов данной длины (по шкале длин). Для ошибки $O2$ дополнительно определялись минимальное и максимальное значение и стандартное отклонение. Результаты тестов для выборок 1 и 2 приведены в таблицах 1 и 2.

Таблица 2. Результаты вычислительного эксперимента 2

N	*K*	*O*1*m*	*O*1*j*	*O2m*	*O2j*	*max O2m*	*max O2j*	*SD O2m*	*SD O2j*
Группа 1									
400	10	10	10	17.60	20.00	24	29	5.22	7.05
600	30	30	30	12.47	14.30	26	26	7.26	5.76
800	42	38	41	9.37	8.12	22	21	5.92	5.33
1000	69	65	68	10.94	10.97	28	26	6.71	5.90
1200	67	64	66	7.25	9.01	22	26	5.22	5.81
1400	126	111	120	6.50	7.70	29	25	5.11	5.78
1600	118	103	109	6.51	6.74	28	27	6.01	5.41
1800	85	70	76	7.43	7.93	30	28	6.86	7.34
Группа 2									
2000	87	70	78	5.66	6.06	21	22	4.90	5.05
4000	97	71	81	4.55	4.60	18	19	3.82	3.40
6000	99	62	73	4.45	4.71	25	24	4.70	4.72
8000	99	50	70	3.66	3.70	14	25	3.03	3.90
10000	99	47	63	3.09	3.22	18	18	2.67	2.39
Группа 3									
30000	46	18	26	2.17	2.69	3	6	0.37	0.82
50000	46	13	21	2.85	3.38	7	12	1.46	2.30
70000	45	19	22	2.11	2.54	3	4	0.31	0.58
90000	46	18	24	2.44	2.71	6	8	0.96	1.21
Группа 4									
100000	8	4	5	3.50	3.60	6	8	1.50	2.24
150000	8	4	5	3.25	3.60	6	8	1.64	2.24
200000	8	3	6	3.67	3.33	6	7	1.70	1.70
250000	8	4	6	2.25	2.67	3	3	0.43	0.47
300000	8	3	5	3.00	4.00	4	9	0.82	2.53
350000	8	3	7	3.67	3.28	6	7	1.70	1.58

В таблицах 1 и 2 столбец *N* содержит значения длин текстов в знаках (без учета пробела), *K* — количество текстов; столбцы *O*1, *O*2 — абсолютные значения ошибок; столбцы *max, SD* — максимальное значение и стандартное отклонение ошибки *O*2. Минимальное значение *O*2 в таблицах 1 и 2 пропущено, так как для всех фрагментов текстов, содержащих больше четырехсот знаков, оно равно двум.

Для наглядного представления о динамике ошибок кусочно-линейная аппроксимация их нормированного значения представлена на соответствующих графиках рисунка 1 и рисунка 2.

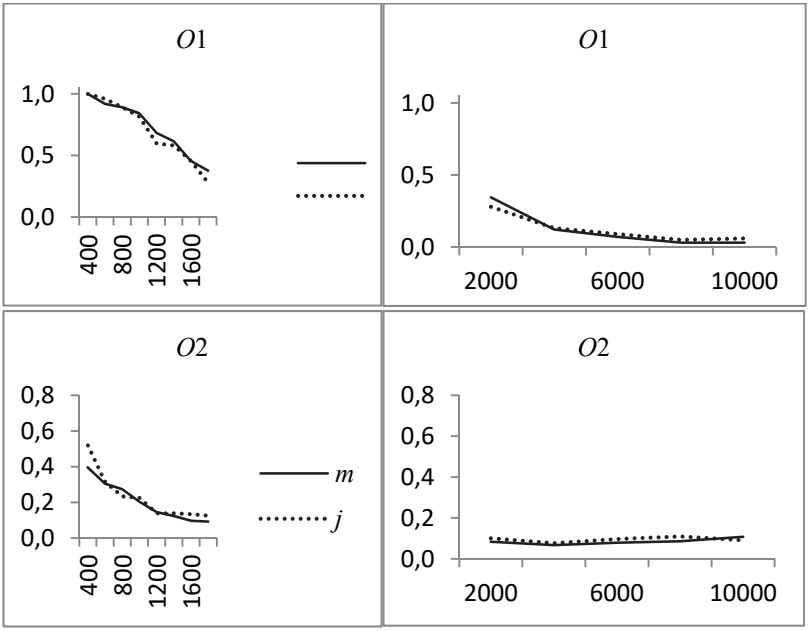

Рис. 1. Графики распределения нормированных ошибок *O*1 и *O*2 методов MAP – *m* и JAC – *j* для групп 1 и 2 фрагментов выборки 1

Индекс *j* в таблицах и на рисунках означает метод Якобсена, который сокращенно обозначим как JAC, индекс *m* — рассмотренный в разделе 4 метод аппроксимации, который сокращенно обозначим как MAP. Оба метода используют одинаковый эталон биграмм [16], независимый от выборок 1 и 2.

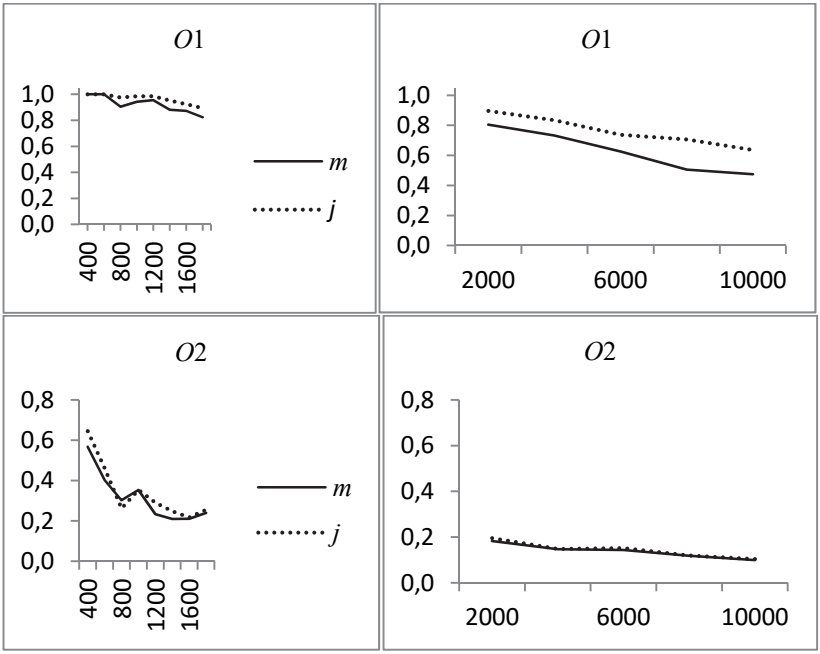

Рис. 2 Графики распределения нормированных ошибок *O*1 и *O*2 методов MAP – *m* и JAC – *j* для групп 1 и 2 фрагментов выборки 2.

Таблица 3. Методы JAC

G	*P*	*P min*	*P max*	*SD P*	*S*	*S min*	*S max*	*SD S*
Выборка 1								
1	34.71	7	82	10.93	1857.38	805	4307	545.76
2	25.09	7	63	8.73	1404.75	678	3060	415.17
3	18.99	8	53	6.91	1135.20	652	2440	294.04
4	16.19	6	41	6.18	1005.70	580	1946	243.17
Выборка 2								
1	43.13	14	97	11.84	2224.71	845	5522	632.14
2	36.55	14	84	11.59	1885.98	836	4017	524.52
3	27.92	10	59	8.33	1504.19	711	2793	404.99
4	23.56	13	35	5.30	1301.48	826	2273	354.28

В таблицах 3 и 4 представлены данные по основным вычислительным характеристикам методов JAC и MAP соответственно. Столбец *G* содержит номер группы текстов, *P* — среднее число выполненных перестановок, столбцы *x min*, *x max*, *SD x* — минимальное, макси-

мальное значение величины x (P, S или U) и стандартное отклонение соответственно.

Таблица 4. Методы MAP

G	*P*	*P min*	*P max*	*SD P*	*U*	*U min*	*U max*	*SD U*
Выборка 1								
1	78.24	15	493	77.22	5.81	1	20	5.23
2	30.59	12	124	11.66	2.28	1	7	0.90
3	22.44	11	48	7.13	1.62	1	4	0.69
4	20.63	11	54	8.12	1.33	1	4	0.72
Выборка 2								
1	119.95	21	572	101.12	8.32	1	20	6.19
2	59.93	19	429	47.57	4.45	1	20	3.50
3	38.31	19	269	19.37	2.90	1	20	1.76
4	36.19	23	63	10.26	2.77	2	4	0.71

Столбец S таблицы 3 содержит среднее общего числа выполненных просмотров в методе JAC, столбец U таблицы 4 — среднее общего числа итераций (частных решений) в методе MAP. Значения P в таблице 4 представляют *сумму* перестановок по всем выполненным итерациям.

6. Обсуждение результатов. Из данных эксперимента можно сделать следующие выводы.

1) Выборки 1 и 2 представляют разные модели текстов. Выборка 1 представляет модель семантически связного, последовательно развивающегося (с точки зрения изложения) текста, словарь которого является наиболее общим для всех носителей языка. Такая модель дает возможность оценить в первую очередь зависимость погрешности метода от объема (длины) одного «усредненного» текста.

Как видно из данных таблицы 1 и графиков рисунка 1, в диапазоне объемов текстов от 400 до 4000 знаков погрешность $O1$ в методе MAP в среднем на 7% больше, чем в методе JAC.

В диапазоне объемов текстов от 4000 знаков и выше погрешность $O1$ в методе MAP в среднем на 26% меньше, чем в методе JAC.

Погрешность $O2$ во всем диапазоне анализируемых текстов в среднем меньше на 17% для метода MAP, чем для метода JAC. Она будет меньше в 20 точках шкалы измерений из 23, чем в методе JAC. То есть при *наличии* ошибки число неправильно идентифицированных знаков для метода MAP будет в среднем меньше.

С учетом указанных различий в целом погрешности методов MAP и JAC зависят от объема «усредненного» текста для текстов выборки 1 приблизительно одинаково.

2) Выборку 2 можно интерпретировать как модель «произвольного» текста. Семантика учебных пособий едина в рамках учебной дисциплины, но различается, иногда значительно, в разных разделах одного пособия. Изложение содержания, как правило, лаконично и ведется с использованием большого числа локальных сокращений. Текст перемежается большим количеством чисел, формул, таблиц и графиков, при формальном исключении которых возникают синтаксические «разрывы». Используются специальные терминологические словари. Каждый случайный фрагмент текста из учебного пособия, в котором оставлены только буквы языка, представляет собой семантический и синтаксический кластер, не всегда и не полностью грамматически правильный и понятный произвольному носителю языка. Усреднение данных по выборке 2 дает возможность оценить зависимость погрешности метода от объема (длины) одного «произвольного» текста.

Данные таблицы 2 и графиков рисунка 2 показывают, что для выборки 2 погрешности $O1$ и $O2$ в методе MAP в среднем на 7% меньше, чем в методе JAC во всем диапазоне шкалы объемов текстов. При этом погрешность $O1$ меньше в 22 точках шкалы из 23, чем в методе JAC. Погрешность $O2$ меньше в 20 точках шкалы из 23, чем в методе JAC.

В целом метод MAP на текстах выборки 2 имеет меньшие погрешности $O1$ и $O2$, чем метод JAC, а их зависимость от объема «произвольного» текста для метода MAP является более стабильной, чем для метода JAC.

3) Метод MAP медленнее, чем метод JAC, что показывают данные таблиц 3 и 4. Сравнительную оценку можно получить исходя из того, что в методе MAP для получения одного частного решения U требуется *не менее* 961 просмотра, аналогичного просмотрам S метода JAC.

Однако из опыта следует, что такое замедление малозаметно и не существенно для решения задачи идентификации знаков текста, особенно с учетом уменьшения погрешности идентификации. При этом из данных таблиц 3 и 4 видно, что уменьшение погрешности в методе MAP обусловлено поиском решения в большем пространстве реальных перестановок, чем в методе JAC.

4) Суммируя данные по выборкам 1 и 2, можно сделать вывод, что погрешность метода MAP в целом меньше, чем метода JAC.

Следует отметить определенную относительность значения ошибки при решении задачи идентификации знаков текста. Например, для одного и того же фрагмента текста получено значение $O2$, равное двум по методу MAP, и четырем — по методу JAC. Однако в первом случае неправильно идентифицированы часто встречающиеся гласные (А, И), тогда как во втором — редко встречающиеся согласные (Г, Б, Ш, Ж).

Какую из двух полученных ошибок — меньшую или большую — будет в данном случае легче исправить, зависит от конкретного текста.

Данные о подобном качественном различии решений, получаемых по методам MAP и JAC, представлены на примере текстов (общее количество — 547) группы 1 выборки 2 в таблице 5. Здесь столбец M представляет метод идентификации, столбцы «ОЕАИ», «НТС» и т.д. — группы идентифицируемых знаков текста, значения столбцов — суммарное количество текстов, в которых встретилась хотя бы одна ошибка при идентификации знаков данной группы.

Таблица 5. Сравнение методов MAP, JAC

M	ОЕАИ	НТС	ВРЛП	КМД	ЙЫЬ	УЯЮЭ	ЗГБЧХ	ШЖЦЩФ
MAP	158	111	178	242	106	179	327	459
JAC	127	94	197	301	147	188	361	500

Из таблицы 5 можно увидеть, что при решении задачи по методу MAP происходит уменьшение ошибки идентификации по сравнению с методом JAC в шести группах знаков из восьми, начиная с третьей, но в первых двух группах погрешность идентификации может возрастать.

7. Заключение. В статье рассмотрены особенности применения методов упорядочения и аппроксимации для решения задачи идентификации знаков текста, и взаимосвязь данных методов. Показано, что метод Якобсена [7] является методом двумерного упорядочивания частот знаковых биграмм, и определены условия, при выполнении которых данный метод имеет наименьшую погрешность при использовании независимого эталона частот буквенных биграмм.

Предложен метод аппроксимации, основанный на сравнении одномерных и двумерных распределений частот знаковых биграмм. При сравнении одномерных условных распределений частот знаковых биграмм используется аппроксимация формы распределений, аналогичная аппроксимации, применяемой в методе Пирсона для проверки однородности двух распределений. Выбор наилучшей аппроксимации осуществляется на основе предложенной оценки погрешности. Частные решения по идентификации знаков текста на основе одномерных условных распределений последовательно улучшаются в ходе итерационного процесса аппроксимации двумерных распределений частот знаковых биграмм.

Это позволяет управлять погрешностью идентификации, которая в целом меньше, чем у метода Якобсена на 12- 17%. Приведены результаты экспериментальной проверки погрешностей предложенного метода аппроксимации и метода Якобсена на двух представительных выборках фрагментов из русскоязычных текстов.

Предложенный в статье метод аппроксимации прост в реализации, является универсальным и может быть использован для идентификации знаков текста любого языка, для которого существует эталонное распределение частот буквенных биграмм.

Литература

1. *Котов Ю.А.* Детерминированная идентификация буквенных биграмм в русскоязычных текстах // Труды СПИИРАН. 2016. Вып. 1(44). С. 181–197.
2. *Шеннон К.* Теория связи в секретных системах // Работы по теории информации и кибернетике // М.: ИЛ. 1963. С. 333–369.
3. *Бабенко Л.К. и др.* Развитие криптографических методов и средств защиты информации // Известия ЮФУ. Технические науки. 2012. № 4. С. 40–50.
4. *Бабенко Л.К., Ищукова Е.А.* Анализ симметричных криптосистем // Известия ЮФУ. Технические науки. 2012. № 12. С. 136–147.
5. *Глухов М.М., Круглов И.А., Пичкур А.Б., Черёмушкин А.В.* Введение в теоретико-числовые методы криптографии // СПб.: Лань. 2011. 400 с.
6. *Минеев М. П., Чубариков В. Н.* Лекции по арифметическим вопросам криптографии // М.: Изд-во «Попечительский совет Механико-математического факультета МГУ им. М.В. Ломоносова». 2010. 186 с.
7. *Jakobsen T.* A fast Method for Cryptanalysis of Substitution Ciphers // Cryptologia. 1995. vol. 19. no. 3. pp. 265–274.
8. *Corlett E.* An Exact A* Method for Solving Letter Substition Ciphers //University of Toronto. 2011. URL: ftp://ftp.cs.toronto.edu/pub/gh/Corlett-MSc-2011.pdf (дата обращения 16.10.2016).
9. *Baragada S.R., Reddy P.S.* A Survey of Cryptanalytic Works Based on Genetic Algorithms // International Journal of Emerging Trends & Technology in Computer Science (IJETTCS). 2013. vol. 2. no. 5. pp. 18–22.
10. *Singh A.P., Pal S.K., Bhatia M.P.S.* The Firefly Algorithm and Application in Cryptanalysis of Monoalphabetic Substitution Ciphers // American Journal of Computer Science and Engineering Survey. 2013. vol. 1. no. 1. pp. 33–52.
11. *Морозенко В.В., Плешкова И.Ю.* О применении генетического алгоритма для криптоанализа шифра Тритемия-Белазо-Виженера // Современные проблемы науки и образования: электронный научный журнал. 2014. № 2. С. 1–11.
12. *Chen J., Rosenthal J.S.* Decrypting classical cipher text using Markov chain Monte Carlo // Statistics and Computing. 2011. vol. 22. no. 2. pp. 397–413.
13. *Bhateja A., Kumar S., Bhateja A.K.* Cryptanalysis of Vigenere Cipher using Particle Swarm Optimization with Markov chain random walk // International Journal on Computer Science and Engineering (IJCSE). 2013. vol. 5. no. 5. pp. 422–429.
14. *Васильев Е.М., Жданова Д.В.* Диахроническое исследование энтропии графем русского письма // Вестник Воронежского государственного технического университета. 2010. № 4. С.1–3.
15. *Васильев Е.М., Гусев К.Ю.* Анализ избыточности русскоязычного текста // Вестник Воронежского государственного технического университета. 2010. № 8. С.1–4.
16. *Жданов О. Н., Куденкова И. А.* Криптоанализ классических шифров // Красноярск: Изд-во Сиб. гос. аэрокосм. ун-та им. акад. М.Ф. Решетнева. 2008. 107 с.
17. *Mohan M., Devi M.K.K., Prakash V.J.* Security Analysis and Modification of Classical Encryption Scheme // Indian Journal of Science and Technology. 2015. vol. 8 no. 8. pp. 542–548.
18. *Губарев В.В.* Введение в теоретическую информатику // Новосибирск: Изд-во НГТУ. 2014. 420 с.

Котов Юрий Алексеевич — к-т физ.-мат. наук, доцент кафедры защиты информации факультета автоматики и вычислительной техники, Новосибирский государственный технический университет (НГТУ). Область научных интересов: информационная и компьютерная безопасность, криптография и криптоанализ, математическое обеспечение вычислительных систем. Число научных публикаций — 25. kotov@corp.nstu.ru; пр. К. Маркса, 20, Новосибирск, 630073; р.т.: +7(383)346-58-03, Факс: +7(383)346-58-03.

DOI 10.15622/SP.50.8

YU.A. KOTOV

APPROXIMATION OF DISTRIBUTIONS OF TEXT CHARACTERS BIGRAMS FREQUENCIES FOR ALPHABETIC CHARACTERS IDENTIFICATION

Kotov Yu.A. **Approximation of Distributions of Text Characters Bigrams Frequencies for Alphabetic Characters Identification.**

Abstract. The article discusses the application features of methods of the frequencies ordering and approximation to solve the problem of text characters identification. The conditions for realization of Jacobsen's method for receiving the least error of identification are defined. The method of approximation of one- and two-dimensional distributions of the frequencies of characters bigrams of the text and the language is offered. The experimental data about errors of Jacobsen's method and the offered approximation method for Russian language texts are provided.

The error of the offered method is less than that of Jacobsen's method. This method can be used for identification of text characters for any language that has a reference distribution of the alphabetic characters bigrams frequencies.

Keywords: approximation, identification, character, bigram, one-to-one substitution, cypher.

Kotov Yuri Alexeevich — Ph.D., associate professor of information protection department of faculty of automation and computer engineering, Novosibirsk State Technical University (NSTU). Research interests: information and computer security, cryptography, software technologies and development of information systems. The number of publications — 25. kotov@corp.nstu.ru; 20, pr. K. Marksa, Novosibirsk, 630073; office phone: +7(383)346-58-03, Fax: +7(383)346-58-03.

References

1. Kotov Yu.A. [Determinate Identification of Russian Text Letter Bigrams]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2016. vol. 1(44). pp. 181–197. (In Russ.).
2. Shennon K. *Raboty po teorii informacii i kibernetike* [Works on the theory of information and cybernetics]. M.: IL. 1963. 832 p. (In Russ.).
3. Babenko L.K. et al. [Development of cryptographic methods and information security tools]. *Izvestija JuFU. Tehnicheskie nauki – Izvestiya SFedU. Engineering sciences*. 2012. vol. 4. pp. 40–50. (In Russ.).
4. Babenko L.K., Ishchukova E.A. [Analysis of symmetric cryptosystems]. *Izvestija JuFU. Tehnicheskie nauki – Izvestiya SFedU. Engineering sciences*. 2012. vol. 12. pp. 136–147. (In Russ.).
5. Gluhov M.M., Kruglov I.A., Pichkur A.B., CHeryomushkin A.V. *Vvedenie v teoretiko-chislovye metody kriptografi* [Introduction to number-theoretic methods in cryptography]. SPb.: Lan'. 2011. 400 p. (In Russ.).
6. Mineev M.P., Chubarikov V.N. *Lekcii po arifmeticheskim voprosam kriptografii* [Lectures on arithmetic cryptography]. M.: Izd-vo «Popechitel'skij sovet Mehaniko-matematicheskogo fakul'teta MGU im. M. V. Lomonosova». 2010. 186 p. (In Russ.).
7. Jakobsen T. A fast Method for Cryptanalysis of Substitution Ciphers. *Cryptologia*. 1995. vol. 19. no. 3. pp. 265–274.
8. Corlett E. An Exact A* Method for Solving Letter Substition Ciphers //University of Toronto, 2011. Available at: ftp://ftp.cs.toronto.edu/pub/gh/Corlett-MSc-2011.pdf (accessed 16.10.2016).

9. Baragada S.R., Reddy P.S. A Survey of Cryptanalytic Works Based on Genetic Algorithms. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*. 2013. vol. 2. no. 5. pp. 18–22.
10. Singh A.P., Pal S.K., Bhatia M.P.S. The Firefly Algorithm and Application in Cryptanalysis of Monoalphabetic Substitution Ciphers. *American Journal of Computer Science and Engineering Survey*. 2013. vol. 1. no. 1. pp. 33–52.
11. Morozenko V.V., Pleshkova I.Yu. [On the application of a genetic algorithm for cryptanalysis of the cipher Triteria-Belazo-Vigenère]. *Sovremennye problemy nauki i obrazovanija: jelektronnyj nauchnyj zhurnal – Modern problems of science and education*. 2014. vol. 2. pp. 1–11. (In Russ.).
12. Chen J., Rosenthal J.S. Decrypting classical cipher text using Markov chain Monte Carlo. *Statistics and Computing*. 2011. vol. 22. no. 2. pp. 397–413.
13. Bhateja A., Kumar S., Bhateja A.K. Cryptanalysis of Vigenere Cipher using Particle Swarm Optimization with Markov chain random walk. *International Journal on Computer Science and Engineering (IJCSE)*. 2013. vol. 5. no. 5. pp. 422–429.
14. Vasil'ev E.M., ZHdanova D.V. [Diachronic study of the entropy of the graphemes of the Russian writing]. *Vestnik Voronezhskogo gosudarstvennogo tekhnicheskogo universiteta – Bulletin of Voronezh State Technical University*. 2010. vol. 4. pp. 1–3.
15. Vasil'ev E.M., Gusev K.Yu. [Redundancy analysis of Russian text]. *Vestnik Voronezhskogo gosudarstvennogo tekhnicheskogo universiteta – Bulletin of Voronezh State Technical University*. 2010. vol. 8. pp. 1–4.
16. Zhdanov O.N., Kudenkova I.A. *Kriptoanaliz klassicheskih shifrov* [Cryptanalysis of classical ciphers]. Krasnojarsk: Izd-vo Sib. gos. ajerokosm. un-ta im. akad. M.F. Reshetneva. 2008. 107 p. (In Russ.).
17. Mohan M., Devi M.K.K., Prakash V.J. Security Analysis and Modification of Classical Encryption Scheme. *Indian Journal of Science and Technology*. 2015. vol. 8 no. 8. pp. 542–548.
18. Gubarev V.V. *Vvedenie v teoreticheskuju informatiku* [Introduction to theoretical informatics]. Novosibirsk: Izd-vo NGTU. 2014. 420 p. (In Russ.).