Методика и результаты сравнительного анализа четырех методов идентификации букв текстов

Ю. А. Котов

Новосибирский государственный технический университет, г. Новосибирск, Россия

Аннотация. В статье приведены результаты сравнения четырех известных частотных методов идентификации букв текстов, необходимые для прикладного решения задач криптоанализа, стеганографии и задач общего анализа текстов, известных в информатике под названием text mining. Для проведения сравнения и получения полной и унифицированной характеристики методов предложена методика оценки, которая включает измерение трех ошибок идентификации и формирование интегральной характеристики на их основе, названной добротностью метода. По данной методике проведено экспериментальное сравнение и качественный анализ одного униграммного и трех биграммных методов идентификации букв текстов. Сравнение выполнено на представительных выборках фрагментов русскоязычных текстов. Определены качественные и количественные особенности методов, границы их эффективного применения, взаимосвязь с типом и объемом обрабатываемого текста.

Ключевые слова: текст, буква, униграмма, биграмма, идентификация, простая замена, шифр, анализ текста.

DOI 10.14357/20718632190304

Введение

Задача идентификации букв текста (ЗИБ) заключается в сопоставлении знака произвольного текста на некотором языке с определенной буквой соответствующего алфавита на основе числовых характеристик, получаемых из данного текста. Решение ЗИБ необходимо при использовании в текстах неопределенных знаковых кодировок и анализе формальных зависимостей отдельных элементов и текстов в целом. Методы решения данной задачи могут быть применены в задачах защиты информации: криптографии, стеганографии, аутентификации текстов и авторов [1-11]; задачах кодирования и мультикодовой коммуникации [1, 3, 12, 13], распознавания знаков и языка сообщения, генерации текстов [13-17], других задачах формального анализа текстов и квантитативной лингвистики и задачах, известных в информатике под названием *text mining* [18-20].

В общем случае решение ЗИБ можно искать как для специально подготовленных (зашифрованных, закрытых, скрытых), так и для открытых текстов, а полученные в ходе решения числовые закономерности использовать для задач сравнения и распознавания текстовой информации. В задачах защиты информации решение ЗИБ обеспечивает: криптоанализ т.н. шифра простой замены; повышение эффективности криптоанализа других шифров – в задачах криптографии; распознавание текста и оцифрованных устных сообщений в произвольных знаковых последовательностях – в задачах стеганографии; лингвистическую аутентификацию текстов и оцифрованных устных сообщений и их авторов; повышение надежности систем автоматического шифрования за счет распознава-

ния и исключения из шифрования специально подобранных и стандартных текстов; коррекцию текстов и оцифрованных устных сообщений в условиях ошибок и помех - в технологиях защиты и аутентификации информации.

Важной группой методов решения задачи идентификации букв текстов является группа частотных методов, использующих частоту появления в текстах знаковых k –грамм [21-23]. Она удовлетворяет двум основным свойствам: вычислительной эффективности и нечувствительности к отсутствию в тексте разделителя слов. Второе из них особенно важно, например, для криптографии, т.к. прием исключения из текста пробелов для затруднения криптоанализа широко известен. При этом очевидно, что погрешность частотных методов имеет статистический характер и обратно пропорциональна объему анализируемых текстов. В соответствии с общими математическими требованиями, необходимыми и достаточными для оценки характеристик методов, в таком случае являются выборочные распределения таких характеристик в зависимости от объемов текстов, включающие средние, минимальные, максимальные значения и их стандартные отклонения, полученные на представительных выборках текстов.

Оценка методов, имеющих статистических характер результатов, на основе выборочных распределений необходима для сравнения и правильного их использования, в том числе для организации формального взаимодействия методов в сложных (интеллектуальных) системах обработки информации. Однако в большинстве случаев в авторских работах либо рассматривается одна какая-либо характеристика [2-19], либо выбирается ограниченный диапазон объемов текста [2-11], либо объемы текста выбираются из одного текста [2-6], либо приводятся одиночные точечные оценки [9, 15-19], либо все это вместе [9, 15]. Субъективизм, неполнота и недостоверность приводимых оценок не позволяют проводить сопоставление, сравнение и анализ большинства подобных методов и рассматривать вопросы их совместного использования.

В целях полной и унифицированной характеристики методов в работе предложена методика оценки и представлены результаты исследования четырех частотных методов идентификации букв текстов: метода простого частотного упорядочивания, основанного на униграммных характеристиках текстов, и методов детерминированной идентификации [24], Якобсена [2, 25], аппроксимации [25], основанных на биграммных характеристиках текстов. Исследования проведены для русскоязычных текстов.

1. Материалы и методы

Для оценки частотных методов идентификации букв текстов недостаточно оценивать метод только по количеству неправильно идентифицируемых букв, как это делается, например, в [2]. Необходимо также учитывать частоту появления таких букв в тексте и частоту точных решений – то есть количества текстов, в которых не выявлено ошибок идентификации. Таким образом, для полной характеристики частотного метода идентификации букв текстов необходимо учитывать три погрешности идентификации. На их основе возможно сформировать интегральную оценку методов.

Будем использовать следующие обозначения.

{{UPO, U}, {JAC, J}, {MAP, M}, {DAT, D}} – сокращенные названия методов, где:

{UPO, U} – метод простого частотного упорядочивания [25];

{JAC, J} – метод Якобсена [2,25];

{MAP, M} – метод аппроксимации [25];

{DAT, D} – метод детерминированной идентификации [24].

x – объем текста в знаках.

O_1 – ошибка, определяемая как отношение количества текстов, в которых обнаружена хотя бы одна ошибка идентификации букв, к общему количеству обработанных текстов; характеризует абсолютную погрешность метода.

O_2 – ошибка, определяемая отношением количества неправильно идентифицированных букв к общему количеству букв в отдельном обрабатываемом тексте; характеризует относительную погрешность метода.

O_3 – ошибка, определяемая отношением суммарного количества появлений в отдельном тексте неправильно идентифицированных букв к общему объему данного текста; характеризует суммарную частотность ошибки O_2, может быть использована для оценки читаемости текста в

Табл. 1. Состав выборок «Тексты 1» и «Тексты 2»

x	K, Текст1	K, Текст2	x	K, Текст1	K, Текст2	x	K, Текст1	K, Текст2
400	20	10	2000	93	87	90000	49	45
600	49	30	4000	99	97	110000	50	45
800	64	42	6000	100	99	150000	19	7
1000	76	69	8000	100	99	200000	20	7
1200	82	67	10000	100	99	250000	20	8
1400	86	126	30000	48	43	300000	20	8
1600	90	118	50000	47	43	350000	19	8
1800	93	85	70000	49	46	Итого	1393	1288

Табл. 2. Временная эффективность методов

Метод	UPO	MAP	DAT	JAC
Время, сек	1	3	5	10

подстановке идентифицированных букв как R=1-O_3.

O_c – ошибка, определяемая для отдельной буквы как отношение количества текстов, в которых буква идентифицирована неправильно, к общему количеству обработанных текстов; характеризует погрешность метода в идентификации отдельной буквы.

$Q = 1 - O_1 \cdot O_2 \cdot O_3$ – интегральная оценка, определяющая общую добротность метода при решении задачи идентификации букв текста.

Данные для сравнительного анализа методов получены в результате вычислительного эксперимента, в котором использовались две выборки фрагментов русскоязычных текстов: «Тексты1» и «Тексты 2», описанные в [25]. Состав выборок приведен в Табл. 1.

На полной выборке текстов 1 (1393 фрагмента) была выполнена оценка времени работы всех четырех методов, данные которой приведены в Табл. 2. Оценка времени выполнена сверху в целых секундах с помощью процедуры *time* языка С на платформе i5-7500 CPU 3.40 GHz/8 GB. Время работы трех методов: UPO, JAC и MAP не зависит от объемов фрагментов, а зависит только от их количества, так как все четыре метода работают не с самим текстом, а его представлением в виде таблиц униграмм и биграмм, т.е. с фиксированным объемом информации. Время на подготовку таблиц зависит от объемов текста, но из оценки исключено. При этом время работы метода DAT зависит от объема текста и может возрастать в два раза при его уменьшении от 10000 до 200 знаков. Это обусловлено тем, что метод DAT представляет собой логико-алгебраический метод [24], и при уменьшении объемов текста количество проверяемых сочетаний биграмм возрастает, т.к. в таблицах биграмм растет количество совпадающих и нулевых значений. При объемах текстов от 4000 знаков и выше метод DAT становится вторым по вычислительной эффективности из четырех рассматриваемых методов после метода UPO.

В Табл. 3 и Табл. 4 приведены средние значения погрешностей O_1 - O_3 и добротности Q исследуемых методов для выборок 1 и 2 соответственно, а также их суммарные значения по всему диапазону измерения («Всего») и при x>2000 («$\sum x$>2000»).

В качестве эталонной упорядоченности по частоте появления букв в русскоязычных текстах была принята последовательность ОЕАИНТСВРЛКДМПУЫЗЯГБЙЬХЧШЖЮЦЩЭФ.

Табл. 3. Средние погрешности методов для Текстов 1 (начало)

Тексты 1	UPO				JAC			
x	O_1	O_2	O_3	Q	O_1	O_2	O_3	Q
400	1	0,853	0,785	0,3302	1	0,519	0,421	0,78133
600	1	0,833	0,756	0,3698	0,960	0,314	0,224	0,93232
800	1	0,811	0,750	0,3921	0,890	0,234	0,151	0,96851
1000	1	0,809	0,739	0,4024	0,820	0,225	0,156	0,97124
1200	1	0,791	0,709	0,4391	0,600	0,14	0,080	0,9933
1400	1	0,778	0,708	0,4492	0,580	0,14	0,070	0,99436
1600	1	0,762	0,664	0,4938	0,460	0,134	0,058	0,99638

Табл. 3. Средние погрешности методов для Текстов 1 (продолжение)

Тексты 1	UPO				JAC			
x	O_1	O_2	O_3	Q	O_1	O_2	O_3	Q
1800	1	0,750	0,661	0,5046	0,280	0,125	0,057	0,99801
2000	1	0,744	0,650	0,5162	0,280	0,100	0,039	0,99889
4000	1	0,727	0,615	0,5529	0,130	0,077	0,016	0,99984
6000	1	0,721	0,609	0,5607	0,090	0,097	0,022	0,99981
8000	1	0,698	0,589	0,5893	0,050	0,11	0,025	0,99986
10000	1	0,697	0,582	0,5938	0,060	0,091	0,019	0,9999
30000	1	0,630	0,523	0,6702	0,020	0,097	0,015	0,99997
50000	1	0,637	0,525	0,6653	0,020	0,064	0,009	0,99999
70000	1	0,639	0,527	0,6631	0,020	0,064	0,009	0,99999
90000	1	0,610	0,494	0,6983	0,020	0,064	0,01	0,99999
110000	1	0,620	0,501	0,6894	0,020	0,064	0,01	0,99999
150000	1	0,604	0,502	0,6968	0,000	0,000	0,000	1
200000	1	0,605	0,494	0,7014	0,050	0,097	0,015	0,99993
250000	1	0,600	0,498	0,7013	0,000	0,000	0,000	1
300000	1	0,593	0,487	0,7107	0,050	0,097	0,015	0,99993
350000	1	0,598	0,489	0,7075	0,000	0,000	0,000	1
Всего:	23	16,111	13,859	13,0982	6,400	2,853	1,422	22,6335
$\sum x$>2000	14	8,980	7,437	9,2007	0,53	0,923	0,164	13,9992
	MAP				DAT			
400	1	0,397	0,291	0,8847	1	0,624	0,529	0,66967
600	0,92	0,305	0,220	0,9383	1	0,459	0,357	0,83582
800	0,89	0,273	0,214	0,9480	0,97	0,357	0,269	0,90688
1000	0,84	0,203	0,152	0,974	0,95	0,284	0,199	0,94637
1200	0,68	0,145	0,094	0,9907	0,84	0,219	0,147	0,97291
1400	0,62	0,124	0,089	0,9932	0,71	0,201	0,142	0,97963
1600	0,46	0,097	0,056	0,9975	0,61	0,162	0,1	0,99012
1800	0,38	0,092	0,042	0,9985	0,53	0,147	0,088	0,99312
2000	0,34	0,084	0,031	0,9991	0,3	0,089	0,025	0,99934
4000	0,12	0,067	0,023	0,9998	0,06	0,075	0,021	0,99990
6000	0,07	0,078	0,038	0,9998	0,02	0,097	0,018	0,99996
8000	0,03	0,086	0,020	0,9999	0	0	0	1
10000	0,03	0,107	0,023	0,9999	0,03	0,086	0,014	0,99996
30000	0,04	0,064	0,010	0,9999	0,02	0,064	0,009	0,99999
50000	0,02	0,064	0,013	0,9999	0	0	0	1
70000	0,02	0,064	0,012	0,9999	0	0	0	1
90000	0	0	0	1	0,02	0,064	0,017	0,99998
110000	0,02	0,064	0,010	0,9999	0,02	0,064	0,016	0,99998
150000	0	0	0	1	0	0	0	1
200000	0	0	0	1	0	0	0	1
250000	0	0	0	1	0	0	0	1
300000	0	0	0	1	0	0	0	1
350000	0	0	0	1	0	0	0	1
Всего:	6,48	2,316	1,339	22,7234	7,08	2,995	1,952	22,2936
$\sum x$>2000	0,35	0,597	0,149	13,9994	0,17	0,452	0,095	13,9998

Табл. 4. Средние погрешности методов для Текстов 2 (начало)

Тексты2	UPO				JAC			
x	O_1	O_2	O_3	Q	O_1	O_2	O_3	Q
400	1	0,897	0,825	0,2600	1	0,645	0,56	0,63882
600	1	0,856	0,796	0,31870	1	0,461	0,352	0,8375
800	1	0,845	0,784	0,33734	0,98	0,262	0,179	0,95401
1000	1	0,855	0,800	0,31543	0,99	0,354	0,263	0,9072
1200	1	0,855	0,799	0,31652	0,99	0,291	0,211	0,9392
1400	1	0,838	0,769	0,35500	0,95	0,248	0,167	0,9606
1600	1	0,829	0,754	0,37534	0,92	0,217	0,133	0,97339

Табл. 4. Средние погрешности методов для Текстов 2 (продолжение)

Тексты2	UPO				JAC			
x	O_1	O_2	O_3	Q	O_1	O_2	O_3	Q
1800	1	0,824	0,746	0,3857	0,89	0,256	0,190	0,95661
2000	1	0,846	0,781	0,33936	0,9	0,196	0,120	0,97886
4000	1	0,808	0,726	0,4127	0,84	0,148	0,074	0,99082
6000	1	0,817	0,739	0,39621	0,74	0,152	0,096	0,98920
8000	1	0,816	0,729	0,40535	0,71	0,119	0,056	0,99522
10000	1	0,797	0,713	0,43176	0,64	0,104	0,036	0,9976
30000	1	0,780	0,662	0,48323	0,6	0,087	0,016	0,99916
50000	1	0,795	0,713	0,43278	0,47	0,102	0,022	0,99895
70000	1	0,773	0,659	0,49096	0,5	0,090	0,021	0,99906
90000	1	0,781	0,674	0,47355	0,51	0,087	0,014	0,99936
110000	1	0,762	0,710	0,45860	0,38	0,087	0,016	0,99946
150000	1	0,770	0,636	0,51031	0,57	0,121	0,02	0,99863
200000	1	0,786	0,720	0,43355	0,71	0,116	0,02	0,99837
250000	1	0,835	0,771	0,35636	0,75	0,086	0,017	0,99893
300000	1	0,794	0,703	0,44154	0,63	0,129	0,02	0,99838
350000	1	0,798	0,714	0,42978	0,88	0,106	0,019	0,99825
Всего:	23	18,759	16,925	9,16013	17,55	4,465	2,622	22,1083
$\sum x$>2000	14	11,114	9,8706	6,15668	8,93	1,535	0,447	13,9614
	MAP				DAT			
400	1	0,568	0,488	0,72285	1	0,681	0,57	0,61213
600	1	0,402	0,312	0,87447	1	0,627	0,530	0,66650
800	0,9	0,302	0,245	0,93342	0,98	0,477	0,357	0,83309
1000	0,94	0,353	0,285	0,90559	1	0,425	0,313	0,86680
1200	0,96	0,234	0,157	0,96479	1	0,39	0,290	0,88686
1400	0,88	0,209	0,154	0,97164	0,99	0,334	0,230	0,92374
1600	0,87	0,210	0,151	0,97238	0,93	0,322	0,23	0,9311
1800	0,82	0,24	0,197	0,96122	0,86	0,3	0,220	0,94316
2000	0,8	0,183	0,127	0,98147	0,94	0,299	0,213	0,93331
4000	0,73	0,147	0,103	0,98899	0,65	0,186	0,121	0,98529
6000	0,63	0,144	0,094	0,99145	0,55	0,154	0,102	0,99135
8000	0,51	0,118	0,078	0,9953	0,43	0,162	0,125	0,99093
10000	0,47	0,099	0,055	0,99740	0,39	0,162	0,117	0,99257
30000	0,37	0,071	0,017	0,99955	0,12	0,161	0,093	0,99821
50000	0,26	0,097	0,021	0,99946	0,19	0,121	0,043	0,99901
70000	0,43	0,073	0,012	0,99961	0,2	0,187	0,17	0,99364
90000	0,38	0,08	0,013	0,9996	0,2	0,150	0,088	0,99735
110000	0,31	0,074	0,014	0,99968	0,16	0,175	0,102	0,99714
150000	0,43	0,13	0,022	0,99877	0,14	0,193	0,036	0,99902
200000	0,43	0,118	0,023	0,99881	0,14	0,387	0,147	0,99202
250000	0,75	0,07	0,012	0,99935	0,13	0,097	0,023	0,99971
300000	0,5	0,105	0,016	0,99916	0,13	0,258	0,037	0,99876
350000	0,5	0,105	0,016	0,99915	0,13	0,387	0,152	0,99237
Всего:	14,87	4,13	2,615	22,2541	12,35	6,731	4,460	21,4979
$\sum x$>2000	6,7	1,428	0,499	13,9663	3,56	2,789	1,357	13,9274

2. Метод простого частотного упорядочивания

Метод идентификации букв текста с помощью частотного упорядочивания знаков [25], называемый далее UPO, считается элементарным и простым и в научных работах в лучшем случае упоминается вскользь. Однако математический смысл метода заключается в приведении текста и эталона к единой системе координат, без чего невозможен числовой анализ знаковых последовательностей, то есть данный метод выполняет функцию преобразования нечисловых координат к единой системе и обеспечивает начальное приближение для частотных методов анализа текстов. Это и определяет

фундаментальное значение и важность метода, а также необходимость изучения его свойств.

Суть метода UPO заключается в подсчете частот встречаемости отдельных знаков в тексте и их упорядочивании по данной частоте, а затем в отождествлении по порядку следования с соответствующими знаками некоторого эталона [25]. Формально метод может быть описан следующим образом. 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. Предположим, что 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)

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

будет минимальной по построению [25]. По результатам одинакового упорядочения знаки анализируемого текста z_i отождествляются с соответствующими знаками эталона e_i , $z_i \equiv e_i$.

Обычно в качестве эталона используются буквы алфавита выбранного языка, упорядоченные в соответствии с т.н. «вероятностью появления» букв в текстах на данном языке [21, 25]. Частота появления знака в тексте является униграммной характеристикой, и по этому критерию данный метод является униграммным.

Так как в качестве эталонов могут использоваться алфавиты разных языков, то метод является универсальным. Методологической основой метода является предположение о том, что при росте объемов исследуемого текста частоты встречаемости отдельных знаков в пределе будут стремиться к «вероятности появления» в тексте отдельных букв. Принято считать, что при неограниченном росте объема анализируемого текста погрешность такого подхода будет стремиться к нулю.

Однако экспериментальные данные (Табл. 3, Табл. 4) показывают, что это не совсем так. При росте объемов текста от 400 до 350000 знаков, то есть почти на три порядка, так и не появилось ни одного текста, в котором бы не было ошибок идентификации, и погрешность O_1 для метода UPO всюду равна единице. Добротность данного метода определяется в основном погрешностью O_2 и вполне представима долей верно идентифицированных букв $y = 1 - O_2(x)$. Графики аппроксимированных значений y для текстов 1 и 2 приведены на Рис. 1 (здесь и далее шкала по оси X неравномерная).

Как видно из данных Табл. 3 и Табл. 4 и графиков на Рис. 1, не менее 77% от максимальной добротности метода достигается при объеме текста около 4000 знаков. При дальнейшем росте объемов текстов добротность возрастает очень

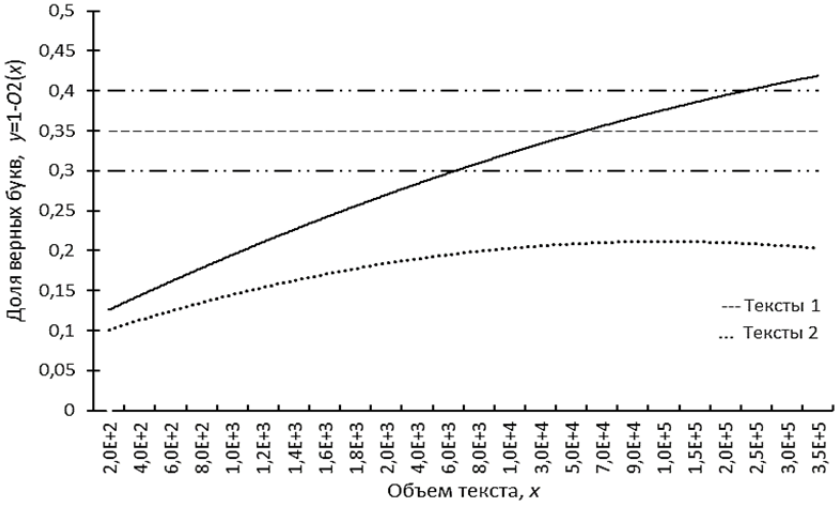

Рис.1. Доля верных букв, получаемых методом UPO

медленно (в лучшем случае – Тексты 1), либо не изменяется (в худшем случае – Тексты 2).

Почему так происходит, показывает поведение ошибки O_c, значения которой для выборок 1 и 3 приведены в Табл. 5. Новая выборка «Тексты 3» была сформирована дополнительно к выборкам 1 и 2. Выборка 3 сформирована случайным образом как по текстам (в нее включены и тексты из выборок 1 и 2, и отсутствующие в них), так и началам фрагментов, и включает в себя 1918 фрагментов: 543 – на интервале объемов текстов 400 – 2000 знаков, 478 – на интервале 2000 - 10000 знаков, 600 – на интервале 10000 – 110000 знаков и 297 – на интервале 100000 - 350000 знаков.

Если проанализировать частоту ошибки идентификации каждой буквы O_c в определенном диапазоне объемов текстов (Табл. 5), то можно увидеть, что при росте объемов текста она значительно снижается лишь для ограниченного числа букв, для других – снижается очень незначительно либо не изменяется. Следовательно, в лучшем случае методом UPO будут правильно идентифицированы 9 - 13 букв текста, при этом наиболее вероятно, что это будут буквы с ошибкой $O_c < 0{,}5$ из левой половины Табл. 6, в которой приведены данные из диапазона 100000 - 350000 Табл. 5, упорядоченные по возрастанию O_c.

То, что добротность метода UPO при $x > 4000$ практически перестает изменяться, показывает и поведение разности ошибок $y=O_2(x) – O_3(x) = Z(x)$ (Рис. 2).

Табл. 5. Побуквенная погрешность метода UPO

Тексты1, буквы	Ошибка O_c в диапазоне объемов текстов:				Тексты3, буквы	Ошибка O_c в диапазоне объемов текстов:			
	400-1800	2000-10000	30000-90000	100000-350000		400-1800	2000-10000	30000-90000	100000-350000
О	0,254	0,065	0,035	0,06	О	0,243	0,125	0,047	0,024
Е	0,617	0,388	0,264	0,376	Е	0,538	0,454	0,280	0,276
А	0,778	0,776	0,690	0,752	А	0,797	0,816	0,718	0,744
И	0,829	0,778	0,700	0,624	И	0,796	0,787	0,658	0,690
Н	0,649	0,537	0,381	0,419	Н	0,709	0,659	0,465	0,444
Т	0,735	0,555	0,442	0,402	Т	0,742	0,625	0,510	0,465
С	0,646	0,441	0,208	0,094	С	0,678	0,527	0,282	0,148
Р	0,785	0,721	0,711	0,761	Р	0,801	0,805	0,673	0,704
В	0,746	0,669	0,629	0,744	В	0,805	0,803	0,718	0,778
Л	0,687	0,547	0,325	0,128	Л	0,768	0,640	0,477	0,391
П	0,829	0,673	0,462	0,333	П	0,845	0,776	0,547	0,485
К	0,788	0,703	0,518	0,419	К	0,801	0,728	0,622	0,576
М	0,858	0,811	0,848	0,709	М	0,867	0,847	0,843	0,852
Д	0,806	0,811	0,751	0,761	Д	0,856	0,824	0,742	0,758
Й	0,888	0,925	0,909	0,949	Й	0,882	0,904	0,940	0,946
Ы	0,891	0,807	0,66	0,632	Ы	0,884	0,77	0,710	0,633
Ь	0,89	0,904	0,924	1,000	Ь	0,921	0,906	0,917	0,912
У	0,870	0,705	0,284	0,077	У	0,867	0,726	0,510	0,387
Я	0,900	0,880	0,777	0,829	Я	0,912	0,862	0,853	0,825
Ю	0,843	0,770	0,579	0,496	Ю	0,831	0,77	0,653	0,623
Э	0,756	0,721	0,589	0,53	Э	0,777	0,753	0,613	0,657
З	0,902	0,911	0,898	0,906	З	0,882	0,893	0,923	0,936
Г	0,884	0,874	0,817	0,889	Г	0,901	0,9	0,865	0,815
Б	0,879	0,858	0,736	0,684	Б	0,884	0,879	0,810	0,815
Ч	0,89	0,919	0,934	0,983	Ч	0,924	0,935	0,947	0,929
Х	0,861	0,868	0,853	0,923	Х	0,877	0,868	0,843	0,788
Ш	0,856	0,854	0,858	0,940	Ш	0,919	0,895	0,937	0,939
Ж	0,845	0,795	0,868	0,940	Ж	0,871	0,837	0,910	0,946
Ц	0,811	0,726	0,584	0,538	Ц	0,853	0,803	0,670	0,680
Щ	0,817	0,756	0,619	0,479	Щ	0,818	0,772	0,635	0,586
Ф	0,651	0,476	0,254	0,367	Ф	0,713	0,605	0,335	0,32

Табл. 6. Группы упорядочивания погрешности O_c метода UPO

Тексты1	O_c	Тексты3	O_c	Тексты1	O_c	Тексты3	O_c
О	0,0598	О	0,0236	Б	0,6838	Р	0,7037
У	0,0769	С	0,1481	М	0,7094	А	0,7441
С	0,0940	Е	0,2761	В	0,7436	Д	0,7576
Л	0,1282	Ф	0,3199	А	0,7521	В	0,7778
П	0,3333	У	0,3872	Р	0,7607	Х	0,7879
Ф	0,3675	Л	0,3906	Д	0,7607	Г	0,8148
Е	0,3761	Н	0,4444	Я	0,8291	Б	0,8148
Т	0,4017	Т	0,4646	Г	0,8889	Я	0,8249
Н	0,4188	П	0,4848	З	0,9060	М	0,8519
К	0,4188	К	0,5758	Х	0,9231	Ь	0,9125
Щ	0,4786	Щ	0,5859	Ш	0,9402	Ч	0,9293
Ю	0,4957	Ю	0,6229	Ж	0,9402	З	0,9360
Э	0,5299	Ы	0,6330	Й	0,9487	Ш	0,9394
Ц	0,5385	Э	0,6566	Ч	0,9829	Й	0,9461
И	0,6239	Ц	0,6801	Ь	1,0000	Ж	0,9461
Ы	0,6325	И	0,6902				

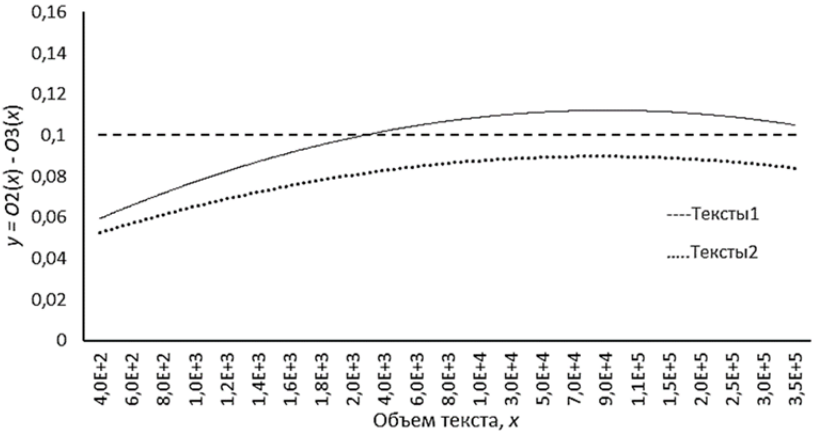

Рис.2. Изменение разности ошибок $y = O_2(x) - O_3(x) = Z(x)$ для метода UPO

Первоначально Z имеет тенденцию к росту до x = 100000 знаков, а затем начинает медленно снижаться. Это говорит о том, что при росте объема текста до 100000 знаков все чаще будут правильно идентифицироваться наиболее частотные буквы из левой половины Табл. 6, а затем и наименее частотные буквы (Ф, Щ и др.). С учетом соотношения приращения Z и объема x при x>4000 знаков можно приблизительно принять Z постоянной величиной, $Z \approx 0{,}1$ с погрешностью ± 0,02 (Рис. 2). Следовательно при $x > 4000$ знаков группа правильно идентифицируемых букв текста стабилизируется по средней частоте встречаемости, что соответствует данным Табл. 5 и Табл. 6. Доля правильно идентифицированных знаков составит при этом 35 ± 5% в лучшем случае (Рис. 1). Так как Z всюду имеет один знак, то можно сделать вывод об устойчивости метода UPO относительно эталонного частотного упорядочивания. Из неравенства $Z > 0$ следует, что средняя частота встречаемости группы правильно идентифицируемых букв сдвинута в область более часто встречающихся букв, чем менее часто встречающихся.

Достигаемое минимальное значение $O_3 = 0{,}49$ говорит о том, что при использовании метода UPO в лучшем случае около половины текста может быть восстановлено правильно и читаемость текста $R = 1 - O_3$ может превышать, пусть и незначительно, половину текста. Эта читаемость такова, что неподготовленный носитель языка сможет, используя только личный тезаурус и автоматизацию перестановок знаков, восстановить остаточный текст приблизительно за 2 – 4 часа и 20 - 40 перестановок пар знаков (данные получены автором на примере выпол-

нения заданий, выданных более чем тремстам студентам в течение четырех лет).

Метод UPO чувствителен к связности текста и наличию в нем лексических погрешностей, что показывает значительное – более чем на 40% – изменение суммарной добротности метода для выборок 1 и 2, для которых она составляет 13,10 и 9,16 соответственно (Табл. 3, Табл. 4 и Рис. 1, Рис. 2). Локальная добротность может изменяться в 1,6 раза (0,7 и 0,43 - Табл. 3, Табл. 4). В целом экспериментальные данные (Табл. 3, Табл. 4 и Рис. 1, Рис. 2) показывают, что добротность метода с ростом объема текста может повышаться, в том числе возможно и после объемов в 350000 знаков. Однако при объеме текстов свыше 4000 знаков этот прирост происходит очень медленно и не во всех случаях. В практических целях можно принять, что приблизительно при объеме текста около 4000 знаков метод достигает или почти достигает своей максимальной добротности. Она характеризуется тем, что доля правильно идентифицированных знаков составляет 35 ± 5% в лучшем случае (Рис.1), а читаемость восстановленного текста приближается к 40 – 50%. Хотя при этом возможна ручная доработка результата, затраты на нее будут достаточно большими, что с учетом чувствительности метода к типу обрабатыываемого текста не позволяет рекомендовать его отдельное прикладное применение. Метод должен использоваться для формирования системы координат и начального приближения в задачах числового анализа знаковых последовательностей.

3. Метод детерминированной идентификации DAT

Входной информацией для метода детерминированной идентификации DAT [24] букв русскоязычных текстов является биграммное представление анализируемого текста, т.е. по данному критерию метод является биграммным. Предполагается упорядоченность этого представления в соответствии с эталонной упорядоченностью. Для работы метода не требуется эталона биграмм, так как метод DAT представляет собой логико-алгебраический метод, основанный на фиксированной последовательности применения системы идентифицирующих отдельные буквы русскоязычных текстов функций **F**, включающей в себя 32 функции, объединенные в 7 групп и формально описанные в [24]. При этом ранее идентифицированные буквы используются при идентификации последующих. Вследствие этого метод не может применяться для текстов на других языках, то есть не является универсальным, а также распространяет ошибку идентификации – ранее неправильно идентифицированная буква может приводить к ошибочной идентификации следующей буквы и так далее. Однако подобное распространение ошибки в методе DAT заметно ограничено (Табл. 3, Табл. 4) и в наибольшей степени этот эффект проявляет себя при уменьшении объемов текста от 4000 знаков (Рис. 3).

Среди других рассматриваемых методов данный метод является единственным, допускающим возможность модификации с целью повышения точности идентификации как отдельных букв, так и букв алфавита в целом. Незначительные изменения всего лишь нескольких функций существенно повысили точность метода по сравнению с первоначальной [24] и приблизили его к методам JAC и MAP (Табл. 3, Табл. 4 и Рис. 3). И хотя данная модификация метода DAT все еще имеет среди рассматриваемых биграммных методов наименьшую суммарную добротность (строка «Всего» Табл. 3, Табл. 4), из данных Табл. 3 (строка «$\sum x$>2000») и графиков на Рис. 3 видно, что для текстов 1 при $x \geq 4000$ добротность метода DAT выше, чем методов MAP и JAC, а для текстов 2 она незначительно меньше. При этом суммарная ошибка O_1 для таких объемов текстов (строка «$\sum x$>2000» Табл. 3, Табл. 4) метода DAT в обоих случаях в 2 – 3 раза меньше, чем методов MAP и JAC, что позволяет говорить о том, что точность метода DAT здесь значительно выше.

Для уточнения этого факта 334 текста, из которых произведены выборки 1 – 3, были разделены на последовательность страниц объемом в 4000, 5000, 6000 и 8000 знаков и для них определена средняя погрешность O_1, получаемая методами DAT, JAC и MAP и соответствующее стандартное отклонение SD O_1 (Табл. 7). Также в Табл. 7 представлены: K_C - общее число страниц текстов, N_A- - количество страниц, в которых используются не все буквы алфавита,

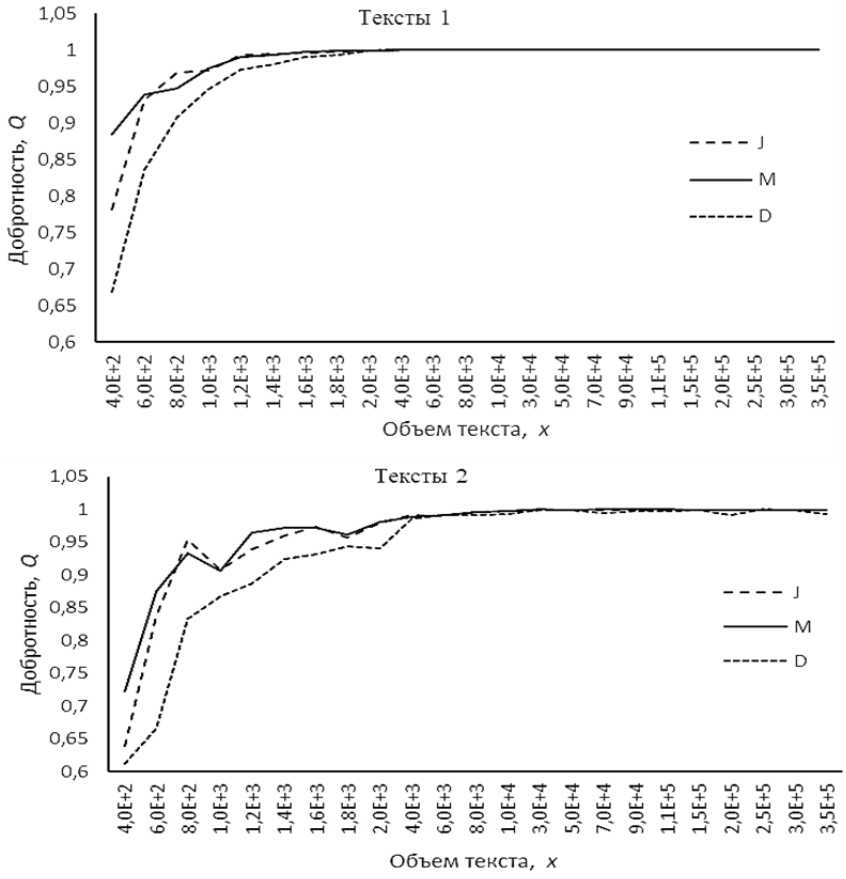

Рис.3. Изменение оценки Q для методов JAC, MAP и DAT

Табл. 7. Связь погрешности O_1 с объемом страниц

x	K_C	N_A-	K_2	DAT			JAC			MAP		
				O_1	SD O_1	O_4	O_1	SD O_1	O_4	O_1	SD O_1	O_4
4000	30091	1017	0,344	0,51	0,22	0,99	0,47	0,3	0,10	0,43	0,26	0,31
5000	24038	522	0,201	0,42	0,23	1	0,42	0,3	0,06	0,38	0,26	0,32
6000	20006	295	0,12	0,37	0,23	1	0,39	0,31	0,06	0,35	0,26	0,29
8000	14967	133	0,05	0,29	0,23	1	0,35	0,32	0,04	0,31	0,28	0,31

K_2 – относительная доля текстов, в которых содержатся 2 или более страниц, использующих не все буквы алфавита, O_4 – ошибка в идентификации пропущенных букв как отношение количества страниц типа N_A- с ошибками к общему числу страниц N_A-. Из данных Табл. 7 следует, что с ростом объемов текста абсолютная погрешность O_1 для метода DAT снижается заметно быстрее, чем для методов JAC и MAP, что наглядно представлено на графиках Рис. 4. Изменение O_1 для метода DAT прямо пропорционально K_2. При этом разница между погрешностями O_1 для методов JAC и MAP практически не изменяется и остается постоянной (Рис. 4).

Более высокий уровень погрешности O_1 для текстов в целом (Табл. 7), чем для выборок из них (Табл. 3, Табл. 4) говорит о чувствительности рассматриваемых методов к полноте используемых букв и другим повторяющимся стилистическим особенностям текста. Наиболее

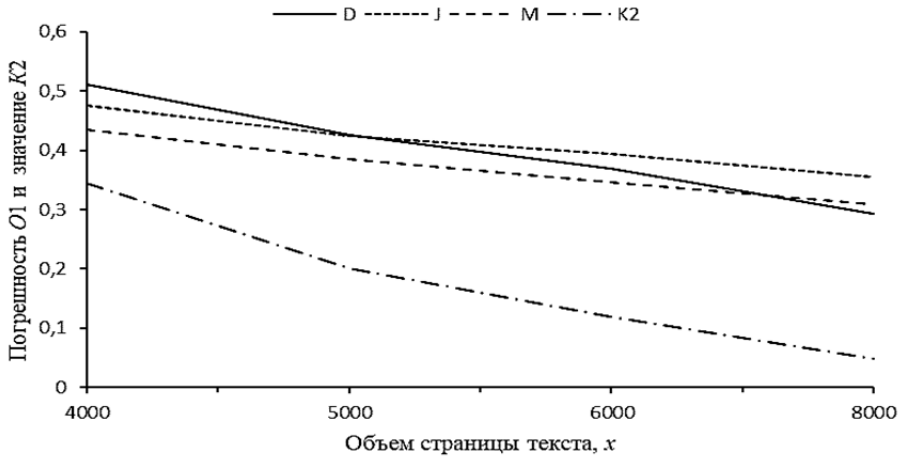

Рис.4. Изменение погрешности O_1 и доли K_2 для методов JAC, MAP и DAT

чувствительным к подобным особенностям текста при объемах текстов $x < 8000$ знаков из трех рассматриваемых методов является метод DAT. Из Табл. 7 видно, что отсутствие в тексте букв алфавита практически в 100% случаев вызывает ошибку при использовании метода DAT и в 30% случаев – при использовании метода MAP, причем независимо от объема обрабатываемого текста (ошибка O_4 Табл. 7).

С другой стороны, для метода JAC ошибка O_4 составляет приблизительно 10% для объемов текста в 4000 знаков, и уменьшается с ростом объема до 3,5% при $x = 8000$ знаков. Отсюда можно сделать вывод, что метод JAC является единственным среди рассматриваемых методов методом, который может идентифицировать пропущенные буквы.

Для оценки влияния на погрешность методов других особенностей текста был проведен следующий эксперимент. Выбрана страница объемом 4000 знаков, на которой не было ошибок идентификации для всех трех методов. Затем на эту страницу было добавлено «слово» длиной в 31 букву, состоящее из перечисления букв алфавита, и удалено соответствующее количество последовательных знаков исходного текста. Затем «слово» удлинялось повторением до тех пор, пока при обработке страницы не появлялась ошибка идентификации. При использовании метода DAT ошибка регистрировалась при размере «слова» 31×16 = 496 знаков, при использовании метода JAC – 31×60 = 1860 знаков, при использовании метода MAP – 31×40 = 1240 знаков. Следовательно, метод JAC наименее зависит от грубых лексических погрешностей в тексте по сравнению с другими методами.

На границу объема текста x ≈ 4000 знаков следует обратить внимание по двум причинам. Во-первых, как показывают все четыре рассматриваемых метода, при таком объеме текста его унарные и бинарные частотные характеристики достигают значений, которые в прикладном плане можно принимать как предельные, в том числе по вероятности. С другой стороны, метод DAT показывает, что при таком объеме текста (если в него не внесены лексические и иные погрешности) появление букв и их сочетаний в нем перестает носить случайный характер, и может быть описано формальным образом. В этом случае появление ошибки идентификации знаков текста будет скорее говорить о погрешности или особенностях в тексте, а не в решении. Из Табл. 3 видно, что первое точное решение получено именно методом DAT, хотя и при $x = 8000$ знаков (для сравнения: методом MAP – при $x = 90000$ знаков, методом JAC – только при $x = 150000$ знаков). Анализ ошибочных фрагментов метода DAT при $x = 4000$ и $x = 6000$ знаков показал, что это именно такие фрагменты, которые содержат лексические погрешности. Следовательно, методы идентификации DAT, JAC и MAP, и особенно DAT, а точнее значения их погрешностей, можно использовать для количественной оценки стилистических особенностей текста.

При этом на практике граница объема текста $x \approx 4000$ знаков указывает тот требуемый максимальный объем текста, которого вполне достаточно для решения с минимальной погрешностью, а во многих случаях и точного решения задач частотной идентификации букв текста, частотного криптоанализа шифра простой замены и других задач, основанных на частотном анализе русскоязычного текста.

4. Методы идентификации JAC и MAP

Метод Якобсена [2] является обобщением метода UPO на двумерный случай - частоты появления в тексте не знаков, а их сочетаний – биграмм. Формально метод заключается в 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, $b_{ij} \in \mathbf{B}$, $t_{ij} \in \mathbf{T}$; **B, T** – таблицы значений частот буквенных биграмм эталонного и анализируемого текста соответственно. Реализация метода представляет собой сортировку по образцу с переменным шагом, адаптированную к решению задачи идентификации знаков текста на основе двумерного распределения частот знаковых биграмм и минимизации суммарной абсолютной разности этого распределения с эталонным распределением [2, 25]. Для работы метода требуется эталон распределения биграмм, то есть он является биграммым методом, а так как может применяться для любых языков, для которых существует такой эталон, – то и универсальным. Суммарная добротность метода JAC – средняя между добротностью методов DAT и MAP (Табл. 3, Табл. 4, Рис. 3). Однако в 11 из 23 точек шкалы измерений (48%) для выборки 1 и в 5 из 23 (22%) – для выборки 2 локальная добротность метода JAC выше, чем метода MAP (Табл. 3, Табл. 4). Это объясняется тем, что погрешность O_c в идентификации наиболее часто встречающихся в текстах шести букв (О, Е, А, И, Т, С) для метода JAC в среднем меньше, а для остальных букв – напротив, больше, чем для метода MAP вплоть до объемов текстов в 100000 знаков (Табл. 8).

Метод MAP, как и метод JAC, является универсальным биграммным методом идентификации букв текстов [25]. В отличие от последнего метод MAP основан на итерационной процедуре аппроксимации распределения биграмм по методу наименьших квадратов по формулам и критериям (3) – (5)

$$s^1_{mj} = \sum_{i=1}^{VB} \frac{(t_{mi} - b_{ji})^2}{t_{mi} + b_{ji}}, \qquad s^2_{mj} = \sum_{i=1}^{VB} \frac{(t_{im} - b_{ij})^2}{t_{im} + b_{ij}} \qquad (3)$$

$$r^1_{mj} = 1 - \frac{s^1_{mj}}{\sum_{i=1}^{VB} b_{ji} + \sum_{i=1}^{VT} t_{mi}}, \; r^2_{mj} = 1 - \frac{s^2_{mj}}{\sum_{i=1}^{VB} b_{ij} + \sum_{i=1}^{VT} t_{im}} \qquad (4)$$

$$W(\mathbf{T}_{n-1}) - W(\mathbf{T}_n) > \varepsilon, \quad \varepsilon = -0.2, \quad \mathbf{T}_0 = \mathbf{T}, \quad n < 20, \qquad (5)$$

и процедуре разрешения коллизий, подробно изложенной в [25].

В оригинальной работе [2] метод JAC назван быстрым методом, но по сравнению с методами DAT и MAP он имеет большее время работы (Табл. 2). В то же время метод JAC является наиболее простым и быстро реализуемым среди них методом. Как уже отмечалось ранее, метод JAC является единственным среди рассматриваемых методов, который может идентифицировать пропущенные буквы и в наименьшей степени зависит от грубых лексических погрешностей в тексте по сравнению с методами DAT и MAP.

Заключение

В работе рассмотрена методика сравнения методов идентификации букв текстов, основанная на получении и сравнении выборочных распределений в зависимости от объемов текстов трех ошибок идентификации: текстовой, знаковой и объемной и связывающей их интегральной характеристики как общей добротности метода. На представительных выборках русскоязычных текстов с использованием данной методики проведен сравнительный анализ четырех известных методов идентификации букв текстов, для которых определены качественные и количественные особенности, оптимальные границы эффективного применения, взаимосвязь с типом и объемом обрабатываемого текста.

Полученные результаты актуальны для решения задач защиты информации: криптографии, стеганографии, аутентификации текстов и

Табл. 8. Погрешность O_c для методов JAC и MAP

Выборка Тексты2, буквы	O_c JAC в диапазоне объемов текстов:				O_c MAP в диапазоне объемов текстов:			
	400-1800	2000-10000	30000-90000	100000-350000	400-1800	2000-10000	30000-90000	100000-350000
О	0,1624	0,0561	0	0	0,2007	0,0624	0,0075	0
Е	0,1204	0,0270	0	0	0,1843	0,0499	0,0075	0
А	0,1551	0,0416	0	0	0,1569	0,0541	0	0
И	0,1697	0,0457	0	0	0,1825	0,0520	0	0
Н	0,1131	0,0353	0	0	0,1004	0,0187	0	0
Т	0,1131	0,0187	0	0	0,1496	0,0665	0,0112	0
С	0,1058	0,0187	0	0	0,1350	0,0665	0,0112	0
Р	0,1715	0,0437	0	0	0,1259	0,0541	0	0
В	0,2500	0,0707	0,0112	0	0,2263	0,0852	0,0075	0
Л	0,1551	0,0229	0,0037	0	0,1058	0,0062	0	0
П	0,2026	0,0520	0	0	0,1460	0,0541	0	0
К	0,2938	0,0603	0	0	0,2391	0,0728	0,0075	0
М	0,4197	0,1476	0,0225	0	0,2774	0,0936	0,0037	0
Д	0,3522	0,1435	0,0150	0	0,2847	0,0811	0,0112	0
Й	0,1843	0,0395	0	0	0,1223	0,0229	0,0037	0
Ы	0,1058	0,0146	0	0	0,0766	0,0083	0	0
Ь	0,0967	0,0166	0,0150	0,0652	0,0766	0,0125	0,0150	0,0870
У	0,0967	0,0166	0	0	0,0858	0,0104	0	0
Я	0,1533	0,0208	0,0037	0	0,1241	0,0104	0	0
Ю	0,1350	0,0187	0,0075	0,0652	0,1296	0,0125	0,0075	0,0652
Э	0,2153	0,0811	0,0262	0,1087	0,2336	0,0582	0,0112	0,0217
З	0,3266	0,1060	0,0075	0	0,2719	0,0790	0,0037	0
Г	0,3631	0,1351	0,0187	0	0,3449	0,1247	0,0225	0,0217
Б	0,4307	0,1559	0,0337	0	0,3631	0,1102	0,0150	0,0217
Ч	0,3522	0,1102	0,0225	0,1304	0,2536	0,0603	0,0187	0,1304
Х	0,3412	0,1247	0,0037	0	0,3029	0,0790	0,0037	0
Ш	0,7719	0,6403	0,4682	0,6522	0,6825	0,4886	0,3071	0,4348
Ж	0,3942	0,1268	0,0150	0,1304	0,3339	0,0977	0,0300	0,1304
Ц	0,5493	0,3888	0,2959	0,4565	0,4653	0,2516	0,1311	0,2391
Щ	0,4252	0,1663	0,0562	0,1087	0,3193	0,0811	0,0412	0,0652
Ф	0,6478	0,4886	0,3858	0,6739	0,6296	0,4262	0,2846	0,4348

их авторов; задач кодирования и мультикодовой коммуникации, распознавания знаков и языка сообщения, генерации текстов, других задач формального анализа и обработки текстов. Они позволяют также провести дальнейшие исследования по снижению погрешности, повышению эффективности и оптимальному применению рассмотренных и аналогичных методов на основе их совместного использования или комплексирования.

Литература

1. Shannon C. Communication theory of secrecy systems // Bell System Technical Journal. 1949. vol. 28. no. 4. pp. 656–715.
2. Jakobsen T. A fast Method for Cryptanalysis of Substitution Ciphers // Cryptologia. 1995. vol.19. no 3. pp. 265-274.
3. Corlett E. An Exact A* Method for Solving Letter Substition Ciphers //University of Toronto. 2011.- ftp://ftp.cs.toronto.edu/pub/gh/Corlett-MSc-2011.pdf.
4. Maya Mohan, M. K. Kavitha Devi, V. Jeevan Prakash Security Analysis and Modification of Classical Encryption Scheme // Indian Journal of Science and Technology. 2015. vol. 8 no. 8. pp. 542–548.
5. Bradly Haner, Ryan Hayward, Grzegorz Kondrak Solving Substitution Ciphers with Combined Language Models // Proceedings of COLING 2014, the 25th International Conference of Computational Linguistics: Technical Papers. Dublin, Ireland, August 23-29. 2014. pp. 2314-2325.
6. Rohit Vobbilisetty, Fabio Di Troia, Richard M. Low, Corrado Aaron Visaggio, Mark Stamp Classic cryptanalysis using hidden Markov models // Criptologia. 2017. vol. 41. no.1. pp.1–28.
7. Bidisha Goswami, Ravichandra G. Public cloud user authentication and data confidentiality using image steganography with hash function // American Journal of Applied Mathematics. 2015. vol.3. no. 1-2. pp. 1-8.

8. James Collins, Sos Agaian High Capacity Image Steganography Using Adjunctive Numerical Representations with Multiple Bit-Plane Decomposition Methods // International Journal on Cryptography and Information Security (IJCIS). 2016. Vol. 6, No. 1-2. pp. 1-21.
9. Воробьева А.А. Методика идентификации интернет-пользователя на основе стилистических и лингвистических характеристик коротких электронных сообщений // Информация и космос. 2017. № 1. С.127-130.
10. Razieh Nokhbeh Zaeem, Monisha Manoharan, Yongpeng Yang, K. Suzanne Barber Modeling and analysis of identity threat behaviors through text mining of identity theft stories // Computers & Security. 2017. no. 65. pp.50-63.
11. Weiming Liang , Haoran Xie, Yanghui Rao , Raymond Y.K. Lau, Fu Lee Wang Universal affective model for Readers' emotion classification over short texts // Expert Systems with Applications. 2018. No. 114. pp. 322—333.
12. Attila Novak, Borbala Siklosi Grapheme-to-Phoneme Transcription in Hungarian // International Journal of Computational Linguistics and Applications. 2016. vol. 7. no. 1, pp. 161—173.
13. Haithem Afli, Loic Barrault, Holger Schwenk OCR Error Correction Using Statistical Machine Translation // International Journal of Computational Linguistics and Applications. 2016. vol. 7. no. 1, pp. 175—191.
14. Grigori Sidorov. Syntactic Dependency Based N-grams in Rule Based Automatic English as Second Language Grammar Correction // International Journal of Computational Linguistics and Applications, Vol. 4, No. 2, pp. 169—188, 2013.
15. Alireza Yousefpour, Roliana Ibrahim, HazaNuzlyAbdel Hamed Ordinal-based and frequency-based integration of feature selection methods for sentiment analysis // Expert Systems with Applications. 2017. no. 75. pp. 80–93.
16. Sanja Štajner, Horacio Saggion, Simone Paolo Ponzetto Improving lexical coverage of text simplification systems for Spanish // Expert Systems with Applications. 2019. no. 118. pp. 80–91.
17. Silvia García-Méndez, Milagros Fernández-Gavilanes, Enrique Costa-Montenegro, Jonathan Juncal-Martínez, F. Javier González-Castaño A library for automatic natural language generation of spanish texts // Expert Systems with Applications. 2019. no. 120. pp. 372–386.
18. Третьяков Ф.И., Серебряная Л.В. Методы автоматического построения рефератов на основе частотного анализа текстов // Доклады Белорусского государственного университета информатики и радиоэлектроники. 2014. № 3. С.40-44.
19. Фомин В.В., Флегонтов А.В., Осочкин А.А. Метод частотно-морфологической классификации текстов // Программные продукты и системы. 2017. №3. С.478-486.
20. Nadir Zanini, Vikas Dhawan Text Mining: An introduction to theory and some applications // A Cambridge Assessment publication. 2015. http://www.cambridgeassessment.org.uk/research-matters/.
21. Абденов А. Ж., Котов Ю. А., Санина О. В. Значения некоторых униграммных характеристик русскоязычных текстов // Научный вестник Новосибирского государственного технического университета. 2017. № 2. С.146-162.
22. Котов Ю. А., Санина О. В. Значения некоторых биграммных характеристик русскояхычных текстов // Вестник СибГУТИ (Сибирский государственный университет телекоммуникации и информатики). 2017. № 4. С.24-34.
23. Котов Ю. А., Санина О. В. Идентификация пробела при неизвестной знаковой кодировке русскоязычных текстов // Вестник СибГУТИ (Сибирский государственный университет телекоммуникации и информатики). 2018. № 4. С.48-60.
24. Котов Ю.А. Детерминированная идентификация буквенных биграмм в русскоязычных текстах // Труды СПИИРАН. 2016. №1. С.181-197.
25. Котов Ю.А. Аппроксимация распределений частот буквенных биграмм текста для идентификации букв // Труды СПИИРАН. 2017. №1. С.190-208.

Котов Юрий Алексеевич. Новосибирский государственный технический университет, г. Новосибирск, Россия. Доцент кафедры защиты информации, канд. физ.-мат. наук, доцент. Количество печатных работ: 37. Область научных интересов: технологии криптографической защиты и аутентификации информации, математическое обеспечение вычислительных машин, комплексов, систем и сетей. E-mail: kotov@corp.nstu.ru

Comparative Analysis of Four Methods for Identifying Letters of Texts

Yu. A. Kotov

Novosibirsk State Technical University, Novosibirsk, Russia

Abstract. The article presents the results of a comparison of four known frequency methods for identifying letters of texts that are necessary for an applied solution of cryptoanalysis, steganography, and general text analysis problems known in computer science as text mining. To compare and obtain a complete and unified characterization of the methods, an evaluation method is proposed, which includes the measurement of three identification errors and the formation of an integral characteristic based on them, called the goodness of the method. According to this method, an experimental comparison and qualitative analysis of one unigram and three bigram methods of identifying letters of texts

was carried out. The comparison was made on representative samples of fragments of Russian texts. The qualitative and quantitative features of the methods, the boundaries of their effective use, the relationship with the type and volume of the text being processed are determined.
It is also shown that an important boundary of text volume for frequency methods and Russian-language texts is a text of approximately 4,000 characters. Such a volume is quite sufficient for the frequency identification of alphabet characters in a Russian-language text with minimal error, and in some cases for obtaining an exact solution. It is shown that with this and a larger amount of text, frequency methods for alphabet characters identification and the proposed estimates of their inaccuracies can be used to quantify certain stylistic features of the text.
Keywords: text, alphabet character, unigram, bigram, identification, one-to-one substitution, cipher, text analysis.

DOI 10.14357/20718632190304

References

1. Shannon C. Communication theory of secrecy systems // Bell System Technical Journal. 1949. vol. 28. no. 4. pp. 656–715.
2. Jakobsen T. A fast Method for Cryptanalysis of Substitution Ciphers // Cryptologia. 1995. vol.19. no 3. pp. 265-274.
3. Corlett E. An Exact A* Method for Solving Letter Substition Ciphers //University of Toronto. 2011.- ftp://ftp.cs.toronto.edu/pub/gh/Corlett-MSc-2011.pdf.
4. Maya Mohan, M. K. Kavitha Devi, V. Jeevan Prakash Security Analysis and Modification of Classical Encryption Scheme // Indian Journal of Science and Technology. 2015. vol. 8 no. 8. pp. 542–548.
5. Bradly Haner, Ryan Hayward, Grzegorz Kondrak Solving Substitution Ciphers with Combined Language Models // Proceedings of COLING 2014, the 25th International Conference of Computational Linguistics: Technical Papers. Dublin, Ireland, August 23-29. 2014. pp. 2314-2325.
6. Rohit Vobbilisetty, Fabio Di Troia, Richard M. Low, Corrado Aaron Visaggio, Mark Stamp Classic cryptanalysis using hidden Markov models // Criptologia. 2017. vol. 41. no.1. pp.1–28.
7. Bidisha Goswami, Ravichandra G. Public cloud user authentication and data confidentiality using image steganography with hash function // American Journal of Applied Mathematics. 2015. vol.3. no. 1-2. pp. 1-8.
8. James Collins, Sos Agaian High Capacity Image Steganography Using Adjunctive Numerical Representations with Multiple Bit-Plane Decomposition Methods // International Journal on Cryptography and Information Security (IJCIS). 2016. Vol. 6, No. 1-2. pp. 1-21.
9. Vorob'eva A.A. Metodika identifikacii internet-pol'zovatelja na osnove stilisticheskih i lingvisticheskih harakteristik korotkih jelektronnyh soobshhenij [The method of identification of the Internet user on the basis of stylistic and linguistic characteristics of short electronic messages]. Informacija i kosmos. 2017. no. 1. pp. 127-130. (In Russ.).
10. Razieh Nokhbeh Zaeem, Monisha Manoharan, Yongpeng Yang, K. Suzanne Barber Modeling and analysis of identity threat behaviors through text mining of identity theft stories // Computers & Security. 2017. no. 65. pp.50-63.
11. Weiming Liang , Haoran Xie, Yanghui Rao , Raymond Y.K. Lau, Fu Lee Wang Universal affective model for Readers' emotion classification over short texts // Expert Systems with Applications. 2018. No. 114. pp. 322—333.
12. Attila Novak, Borbala Siklosi Grapheme-to-Phoneme Transcription in Hungarian // International Journal of Computational Linguistics and Applications. 2016. vol. 7. no. 1, pp. 161—173.
13. Haithem Afli, Loic Barrault, Holger Schwenk OCR Error Correction Using Statistical Machine Translation // International Journal of Computational Linguistics and Applications. 2016. vol. 7. no. 1, pp. 175—191.
14. Grigori Sidorov. Syntactic Dependency Based N-grams in Rule Based Automatic English as Second Language Grammar Correction // International Journal of Computational Linguistics and Applications, Vol. 4, No. 2, pp. 169—188, 2013.
15. Alireza Yousefpour, Roliana Ibrahim, HazaNuzlyAbdel Hamed Ordinal-based and frequency-based integration of feature selection methods for sentiment analysis // Expert Systems with Applications. 2017. no. 75. pp. 80–93.
16. Sanja Štajner, Horacio Saggion, Simone Paolo Ponzetto Improving lexical coverage of text simplification systems for Spanish // Expert Systems with Applications. 2019. no. 118. pp. 80–91.
17. Silvia García-Méndez, Milagros Fernández-Gavilanes, Enrique Costa-Montenegro, Jonathan Juncal-Martínez, F. Javier González-Castaño A library for automatic natural language generation of spanish texts // Expert Systems with Applications. 2019. no. 120. pp. 372–386.
18. Tret'jakov F.I., Serebrjanaja L.V. Metody avtomatich-eskogo postroenija referatov na osnove chastotnogo analiza tekstov [Methods of automatic construction of abstracts based on frequency analysis of texts]. Doklady Belorussskogo gosudarstvennogo universiteta informatiki i radiojelektroniki. 2014. no. 3. pp.40-44. (In Russ.).
19. Fomin V.V., Flegontov A.V., Osochkin A.A. Metod chastotno-morfologicheskoj klassifikacii tekstov [Method of frequency-morphological classification of texts]. Programmnye produkty i sistemy. 2017. no.3. pp.478-486. (In Russ.).

20. Nadir Zanini, Vikas Dhawan Text Mining: An introduction to theory and some applications // A Cambridge Assessment publication. 2015. http://www.cambridgeassessment.org.uk/research-matters/.
21. Abdenov A.J., Kotov Yu.A., Sanina O.V. [Values of some unigram characteristics of Russian texts]. Nauchnyj vestnik Novosibirskogo gosudarstvennogo tehnicheskogo universiteta. 2017. № 2. pp.146-162. (In Russ.).
22. Kotov Yu.A., Sanina O.V. [Values of some bigram characteristics of Russian texts]. Vestnik SibGUTI (Sibirskij gosudarstvennyj universitet telekommunikacii i informatiki). 2017. № 4. pp.24-34. (In Russ.).
23. Kotov Yu.A., Sanina O.V. [Space identification with unknown sign encoding of Russian texts]. Vestnik SibGUTI (Sibirskij gosudarstvennyj universitet telekommunikacii i informatiki). 2018. № 4. pp.48-60. (In Russ.).
24. Kotov Yu.A. [Determinate Identification of Russian Text Letter Bigrams]. SPIIRAS Proceedings. 2016. no 1. pp.181-197. (In Russ.).
25. Kotov Yu.A. [Approximation of Distributions of Text Characters Bigrams Frequencies for Alphabetic Characters Identification]. SPIIRAS Proceedings. 2017. no 1. pp.190-208. (In Russ.).

Kotov Y. A.. Ph. D. (Phys. & Math), associate professor at the department of Information Security, the faculty of Automation and Computer Engineering, Novosibirsk State Technical University, 20 K.Marx av., Novosibirsk, 630073, Russia, e-mail: kotov@corp.nstu.ru