ISSN 0869-8325

ТОМ 16 Выпуск 1

1 – 6 VI • 2008

5 – 11 X • 2008

ОБОЗРЕНИЕ ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ МАТЕМАТИКИ

В выпуске:

Секция **«Прикладная вероятность и статистика»**
Секция **«Дискретная математика»**

Седьмая Международная Петрозаводская конференция
«Вероятностные методы в дискретной математике»
Научные доклады. Часть III

Пятнадцатая Всероссийская школа-коллоквиум
по стохастическим методам

Девятый Всероссийский симпозиум
по прикладной и промышленной математике
Осенняя сессия. Научные доклады. Часть II

Редакция журнала «ОПиПМ» • МОСКВА
2009

ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 16 МАТЕМАТИКИ Выпуск 1
2009

МАЛЮТОВ М. Б., БРОДСКИЙ С.

MDL-ПРОЦЕДУРА ДЛЯ АТРИБУЦИИ АВТОРСТВА ТЕКСТОВ

Рассматривается новый стилометрический атрибутор, не зависящий от контекста: кусочная условная сложность сжатия (*CCC*) литературных текстов. Подход подсказан колмогоровской невычисляемой условной сложностью. Там, где другие методы стилометрии могут иногда не различить похожих авторов, *CCC*-атрибутор, впервые введенный в [12], асимптотически минимален для истинного автора, если выборочные и изучаемые тексты являются достаточно большими, сжатие — достаточно хорошим и выборочное смещение отсутствует. Этот классификатор является упрощением введенного в [20] критерия однородности (частично основанного на сжатии) при естественном ограничении на безусловные средние сложности изучаемых текстов. Он состоятелен при аппроксимации большого текста как стационарной эргодической последовательности, что следует из нижней минимаксной границы сжатия кусочно стационарных строк [17]. Надлежащие параметры нашего классификатора определены в работе [13] по атрибуции «федералистских статей» (Мэдисон против Гамильтона), давшей те же результаты, что и предшествовавшие атрибуторы. Мы также установили там состоятельность *CCC*-атрибутора для *IID*-последовательностей элементарными комбинаторными аргументами и моделированием и использовали эмпирическую *CCC*-асимптотическую нормальность (следующую из [22]) для аппроксимации *P*-значения нашего вывода. Работа [13] содержит также некоторые *новые* интригующие результаты по атрибуции текстов.

Здесь мы обсуждаем дальнейшие свойства атрибуторов на основе сжатия и используем *CCC*-атрибутор для анализа раннего (1925 г.) рассказа М. Шолохова. Показана существенная разница в стиле двух частей этого рассказа.

Ключевые слова: сложность сжатия, атрибуция авторства.

§ 1. Дискриминация посредством универсального сжатия

К. Э. Шеннон [21] создал теорию передачи информации, основанную на статистической теории Колмогорова. В частности, при заданном распределении на алфавите средняя длина сжатия Шеннона–Фано *IID*-строк с элементами из этого алфавита асимптотически достигает своей нижней границы — энтропии Шеннона для длины (сложности) сжатия на символ строки. А. Н. Колмогоров [9] развил теорию сложности *индивидуальных* строк, согласно которой для строк, принадлежащих статистической совокупности, их средняя сложность аппроксимирует их эн-

тропию. Эта идея почти немедленно использовалась Б. М. Фитингофом [24, 25] и позже автором Davisson для того, чтобы создать так называемое *универсальное сжатие* (*UC*), которое, адаптируясь к *неизвестному* стационарному эргодическому распределению (*SED*) строк, достигает асимптотически нижней энтропийной границы. $\mathcal{P}$ — это класс *SED* источников, аппроксимируемый n-цепями Маркова (n-*MC*). Сжатие семейства $\mathbf{L} = \{L_n\colon\ \mathbf{B}^n \to \mathbf{B}^\infty,\ n = 1, 2, \ldots\}$ называется универсальным, если для любых $P \in \mathcal{P}$ и $\varepsilon > 0$, $\mathbf{B} = \{0, 1\}$

$$\lim_{n\to\infty} \mathbf{P}\{P(x) \in \mathbf{B}^n :\ |L_n(x)| + \log P(x) \leqslant n\varepsilon\} = 1,$$

где $|L(x)|$ есть длина $L(x)$, и $|L_n(x)| + \log P(x)$ называется *индивидуальной избыточностью*.

Таким образом, для строки, генерируемой *SED*, асимптотически длина *UC*-сжатия — ее отрицательный логарифм правдоподобия. Это является чрезвычайно плодотворным результатом, если правдоподобие не может быть оценено аналитически. Сначала *UC* использовали для оценки параметров аппроксимации n-*MC* для того, чтобы привести к хорошему сжатию. Потребовалось более десяти лет для того, чтобы появился гораздо более тонкий метод, осуществляющий достаточно полно вышеупомянутую идею Колмогорова для компрессоров. Это было достигнуто в двух *UC* авторами Lempel и Ziv (*LZ*-1977-78).

Оба *LZ*-сжатия не используют никакой статистики строк вообще. Вместо этого *LZ*-78 последовательно строит бинарное дерево *образов*, не встречавшихся до этого в строке, начиная с первого символа строки. Wyner и Ziv доказали, что *LZ*-78 есть *UC*, из чего вытекает следующее:

$$\lim_{n\to\infty} \mathbf{P}\{|L_n(x)|/|x| \to h\} = 1 \quad \text{при} \quad |x| \to \infty$$

для $P \in \mathcal{P}$, где h — двоичная энтропия (на один символ), которая представляет собой нижнюю границу для сжатия *SED* у Шеннона [21]; там *SED* строки впервые введены им как удобные модели естественного языка. К концу восьмидесятых версии *LZ*-78 стали каждодневными инструментами в компьютерной практике. Публикация Риссаненом принципа *минимальной длины описания* (*MDL*) в 1978 г. (продолженная в его статье [18]), и работа Ziv [23] инициировали применения *UC* к статистическим проблемам для *SED* источников. Это было продолжено в работах Б. Рябко с соавторами. Особый интерес для наших применений представляет критерий однородности из [20].

§ 2. Критерий однородности с использованием *CC*

Обозначим $|A|$ и $|A_c|$ соответственно длины бинарной строки A и ее сжатия A_c. Для их отношений введем обозначение $CC_r = |A_c|/|A|$.

Пополненная строка $S = AB$ начинается с A вплоть до текста B без перерыва.

Статистический критерий однородности двух бинарных строк Рябко–Астола строится по

$$T = h_n^*(S) - |A_c| - |Q_c|,$$

где эмпирическая шенноновская энтропия h_n^* пополненной строки S (основанной на n-MC аппроксимации) есть формула (6) в работе [20] упомянутых авторов. Локальная не зависящая от контекста структура (микростиль) длинного (несколько килобайт) литературного текста (ЛТ) может быть достаточно точно смоделирована посредством одной бинарной последовательности n-MC с n не меньшим, чем несколько дюжин. Таким образом, ее вычисление для ЛТ очень сложно в вычислительном отношении и нестабильно для текстов среднего размера, требующих регуляризации малых или нулевых оценок для переходных вероятностей. Поэтому ее применение для ЛТ вместо одинаково трудоемкой в вычислительном отношении процедуры из [19], основанной на асимптотически оптимальном методе максимального правдоподобия и n-MC обучении, не слишком оправданно. Для короткого ЛТ точность SED модели может быть недостаточной, в то время как для очень большого ЛТ (такого, как повесть) из-за соотношений литературной формы («архитектуры») микростиль описывает только локальную часть авторского стиля, как подчеркнуто в работе [4].

Условная сложность сжатия и более наглядная относительная условная сложность сжатия, $0 < CCC_r < 1$, текста B для заданного текста A определяются, соответственно,

$$CCC(B|A) = |S_c| - |A_c|, \quad CCC_r(B|A) = CCC(B|A)/|B|.$$

В наших обозначениях CCC похож на более абстрактную условную сложность Колмогорова и измеряет, как адаптация к *образам* обучающего текста помогает сжать изучаемый текст. CCC аппроксимирует наиболее мощный критерий однородности, основанный на отношении максимального правдоподобия Q, A при наших условиях на размер выборки и состоятельности SED аппроксимации для обоих Q, A.

В нашем исследовании мы усредняем CCC_r кусков одинаковой длины L: Q_i, $i = 1, \ldots, m = [|Q|/L]$ *изучаемого текста* для заданного надежно атрибутированного текста A. Применяемое универсальное сжатие UC одинаково для всех кусков текста:

$$\overline{CCC_r(Q|A)} := \sum_{i=1}^{m} \frac{CCC_r(Q_i|A)}{m}, \quad \overline{CC_r(Q)} := \sum_{i=1}^{m} \frac{CC_r(Q_i)}{m}.$$

Будем называть эти две последние эмпирические величины *средним* $CCC_r(Q)$ и *средним* $CC_r(Q)$ соответственно. Рассмотрим $U(Q, A) = CCC(Q|A) - |Q|$. Если $\overline{CC_r(Q)} \geqslant \overline{CC_r(Q')}$ и $CCC(Q|A) < CCC(Q'|A)$ значимы, то $U(Q, A) > U(Q', A)$ асимптотически для больших выборок.

Величина $U(Q,A)$ похожа на T-критерий однородности Рябко–Астола. В $U(Q,A)$ мы *заменяем* их эмпирическую энтропию Шеннона h^* пополненной выборки S (основанной на n-MC аппроксимации) на $|S_c|$, поскольку обе величины асимптотически эквивалентны $h(|Q|+|A|)$ для идентичного распределения Q, A с энтропией h и превышают это количество для различных распределений A, Q (состоятельность). Критерий T асимптотически инвариантен при перестановке A, Q и строго положителен для различных A, Q, если $a<|A|/|Q|<1/a$, $a>0$. Последнее (но не первое) свойство справедливо также для $U(Q,A)$ в некотором диапазоне $|A|/|Q|$, что следует из нижней минимаксной границы сжатия кусочно стационарных строк [17], логарифмической от $(|Q|+|A|)$.

Для нашего приложения $|A|/|Q|$ велики для того, чтобы статистически оценивать значимость различий между средними $CCC_\tau(Q)$, и $|Q|\geqslant$ 2000 байтов (приблизительно) для применимости *SED* аппроксимации. В [13] показано, что CCC-атрибуция имеет хорошую дискриминирующую способность в этом диапазоне для умеренных величин Q. Кусочный CCC-атрибутор может быть заменен на нашу статистику однородности $U(Q,A)$ в удивительно широком диапазоне случаев с *незначимо изменяющейся средней безусловной сложностью сжатия*, которую легко проверить, используя асимптотическую нормальность CC [22]. Ее весьма правдоподобное обобщение для CCC теоретически оправдывает довольно необычное соотношение между размерами выборки для UC-атрибуции авторства: *один* из *размеров выборки должен значительно превосходить тот*, независимые копии которой изучаются. Например, если мы фиксируем обучающий текст A и сравниваем независимые копии Q_i, $i=1,\ldots,N$, изучаемого текста Q, то $\mathbf{D}\,CCC(Q_i|A)$ имеет порядок $|Q|$, в то время как среднее $CCC(Q|A)$ для различных распределений Q и A превышает те же для идентичных распределений на величину порядка $o(|A||Q|^b)$ при любом $b>0$ (точная верхняя граница отсутствует пока даже для *LZ*-78, нижняя граница, согласно [17], есть $\log(|A||Q|)$. Таким образом, t-отношение незначимо в асимптотике $|A|\to\infty$, $0<\varepsilon<|Q|/|A|$. В [12] это неформально объясняется таким образом: если размеры обучающего текста A и изучаемого текста альтернативного стиля Q сравнимы, то UC адаптируется к распределению пополненной строки S за $o(|A||Q|^b)$ шагов для любого $b>0$, и смещение в $CCC(Q|A)$ поглощается шумом с дисперсией $\mathbf{D}\,CCC(|A||Q|)$ порядка $|(A|Q)|$. Это делает требования [5] (будем называть их ниже CV) сравнимости размеров выборки и симметрии расстояния (вместо критически важной статистической устойчивости выводов) сомнительными и объясняет примеры неправильной классификации по CV в работе [26]. Это может также объяснить корни ранней горячей дискуссии работы [2], где вопрос о размере выборки даже не поднимался.

Мы сохраняем наш CCC-атрибутор (а не $U(Q,A)$) как reference test из-за близкого отношения к первоначальной колмогоровской идее.

§ 3. Сжатый обзор методов микростилометрии

Мы имеем дело только с критериями, не зависящими от содержания текста, а также оставляем в стороне методы, основанные на грамматике. Независимая от содержания атрибуция одинаково применима к любому языку, даже к шифрованным сообщениям, которые еще не прочитаны (таким, как недавно обнаруженные кумранские рукописи). Однако эти методы не всегда робастны по отношению к грамматическим ошибкам, и их дискриминирующая способность может быть ниже семантических атрибуторов.

Одним из препятствий для применения этих методов является тот факт, что стиль писателя со временем меняется (обогащается). Таким образом, мы можем только сравнивать тексты, написанные примерно в одно и то же время.

Также следует заметить, что авторы могут работать в различных литературных формах (например, проза, стихи), которые могут иметь различные статистические свойства. Таким образом, необходимы предварительная подготовка всего текста и разделение его на однородные части. Опечатки должны быть исправлены, а имена собственные — удалены. Следует также заметить, что текст с дополнительной информацией может быть более полезным для компьютерного анализа. Такой дополнительной информацией могут быть ударения в стихах или указание на время, когда текст был создан.

Пионерские работы [15], [16] по стилометрии были основаны на гистограммах распределения длин слов различных авторов и вычислялись по пяти отрывкам текстов длиной 1000 слов для каждого автора. Эти работы показывают значимое различие этих гистограмм для различных авторов (Чарльз Диккенс против Джона Стьарта Милля), писавших на одном языке. В то же время гистограммы для работ Диккенса были близки к гистограммам работ Теккерея в смысле их статистического разброса, вычисляемого по повторным выборкам. Обзоры по стилометрии (включающие в себя применения нашего нового CCC-атрибутора) можно найти в работах [12] и [13]. Эти обзоры не покрывают последние приложения оценки параметров законов Ципфа и Хипса (см., например, работы [14], [8]), основанные на частотах слов, а не на частотах произвольных частей текста, как это делается в нашем атрибуторе.

Неформальное описание в [6] почти двадцатилетнего изучения нескольких сотен атрибуторов студентами Claremont McKenna College, Калифорния, США, показывает, что предварительная обработка текстов (включающая в себя удаление имен собственных и разделение текста на однородные части) не производилась; статистическая стабильность, эволюция стиля, оценка значимости множественных решений и другие важные аспекты не были приняты во внимание. Таким образом, сомнительный уровень руководства работами не позволяет оценить

доброкачественность выводов.

Наше приложение полезно сопоставить с обстоятельной работой [7].

Опуская обсуждение других популярных атрибуторов, мы перейдем прямо к CCC-атрибутору, который показывает даже лучшие характеристики в некоторых приложениях (см., например, работу [10]) до их настройки и улучшений в работе [13]. В [5] приведен обзор многочисленных предшествующих подходов к рассматриваемой проблеме: классификация и кластеризация текстовых библиотек сравнимых размеров с использованием «метрики подобия», напоминающей информационные расстояния Беннетта и др., и под воздействием колмогоровской сложности (КС) [9] с помощью замены КС на практическое универсальное сжатие, удовлетворяющее некоторым свойствам. Симметрия расстояния представляла некоторую проблему в этих статьях в противоположность нашему подходу (см. [11], [2]). Главное отличие нашего метода от всех предыдущих подходов состоит в использовании многих частей изучаемого текста, позволяющем провести статистический анализ их условных сложностей: оценить центр и разброс их распределений. Таким образом, мы можем судить о статистической значимости CCC-разностей аналогично [15], [16]. Мы показали в работе [13], что известный NCD-атрибутор Ли и др. не позволяет атрибутировать авторство в случаях, которые дискриминируются с высоким уровнем значимости нашим CCC-атрибутором.

Так как распределение CCC по кускам не известно, мы иллюстрировали наши результаты в работе [13], главным образом, посредством (довольно убедительных) гистограмм и графиков, показывающих достаточность и приближенную нормальность CCC-атрибутора. Последнее использовалось также для вычисления P-значения атрибуции.

§ 4. Методология

Надежно атрибутированный текст называется *обучающим*. Сжатие и текст, который мы исследуем, называются *изучаемыми*. Изучаемые тексты могут исследоваться многократно для изучения свойств атрибуторов.

Обычно мы выбираем куски Q_i равной длины у изучаемых и обучающих текстов Q и $T(k)$ и вычисляем средние $CC_\tau T(k)$, $CCC_\tau(Q|\,T(k))$, а также их эмпирические среднеквадратичные отклонения для каждого изучаемого и обучающего текста $T(k)$, чтобы облегчить их статистический анализ. Хотя CCC_τ несимметрична, ее можно трактовать как *обобщенное расстояние*. Разбиение текстов на равные куски, как было установлено эмпирически, не влияет на свойства нашего атрибутора, сравнительно с большей вычислительной трудоемкостью при смещении начала кусков для того, чтобы включить все слово целиком. Таким образом, мы предполагаем, что *образы среднего размера наиболее су-*

щественны для CCC-различения, в отличие от таких различных универсальных сжатий, как WinZip, BWT и др., что указывает на еще одно свойство инвариантности нашего метода.

Если среднее $CC_r(Q)$ отличается значимо от среднего $CC_r(T)$ (что может быть установлено с использованием их асимптотической нормальности в [22]), автор обучающего текста T вряд ли будет автором текста Q. Если средние CC_r от Q и T не различаются значимо, то чем меньше среднее $CCC_r(Q|T)$, тем очевиднее совпадение стилей двух текстов. Мы ожидаем, что среднее CCC_r будет наименьшим при обучении на текстах их автора. Предположение о том, что безусловная сложность изучаемого и обучающего текстов приблизительно равны, иллюстрируется в экстремальном случае длинного изучаемого текста, состоящего из повторяющегося символа. Его CCC_r — наименьшая для всех обучающих текстов.

Мы сравнивали CCC_r для различных изучаемых текстов, написанных примерно в одно и то же время, используя t-критерий и непараметрический критерий Вилкоксона.

§ 5. Разница в стиле частей раннего рассказа М. Шолохова

Шолоховский рассказ «Путь-дороженька» был опубликован впервые в период с 25.04.1925 по 21.05.1925 в московской газете «Юный Ленинец» (выпуски 93–97, 99, 101–104, 106–114). Шолохову в это время было 20 лет. За его плечами были четыре года начальной сельской школы во время гражданской войны. Потом — короткое пребывание в тюрьме по обвинению в коррупции во время его службы в продотряде. Он покинул Дон и приехал в Москву в конце 1922 г. Сменил несколько мест работы и опубликовал свой первый короткий рассказ в 1924 г. Какие-то изменения в его писательском стиле между первой и второй частями рассказа «Путь-дороженька» представляются маловероятными.

После предварительной обработки текста (включающей в себя, в частности, удаление имен собственных) мы разбили каждую часть рассказа на 30 равных частей по 2000 байт каждая. Средние безусловные сложности ($\overline{CC}$) были статистически одинаковы. Средняя (intra)-CCC (CCC кусков, обучаемой на остающемся тексте той же части) сравнивалась со средним inter-CCC каждого куска, обучаемой на другой части. Их среднеквадратичные отклонения статистически не различались. Разность между средней inter-CCC и средней intra-CCC оказалась значимой, превышая в четыре раза ее среднеквадратичное отклонение.

Детали вычислений таковы: мы посчитали 30 inter-CCC (кусок Части 2|Часть 1 целиком) и 30 intra-CCC (кусок Части 1|остающийся текст Части 1). Средняя inter-CCC: $M_1 = 576,78$; средняя intra-CCC: $M_2 = 559,43$, их разность равна 17,34, среднеквадратичное отклонение (Mean inter-CCC) равно $s_1 = 2,49$, среднеквадратичное отклонение

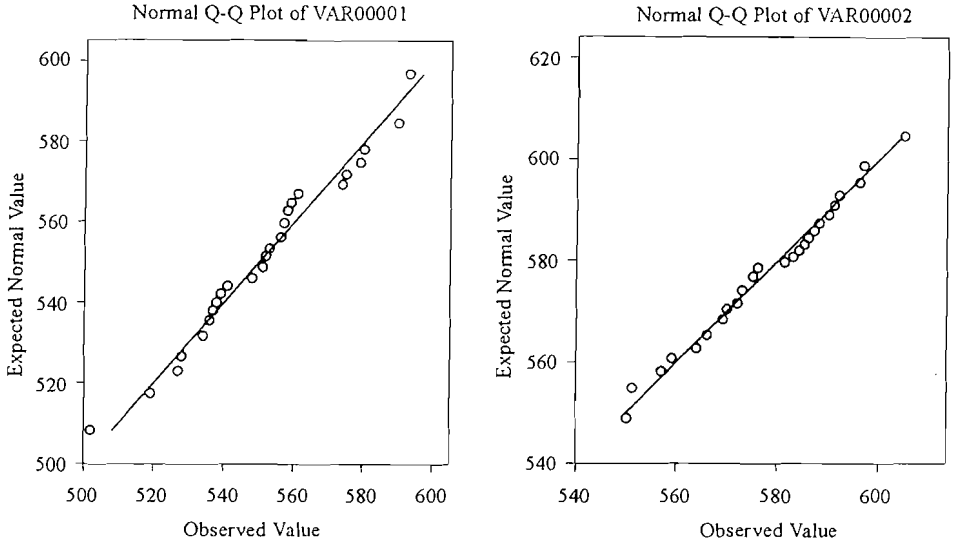
(Mean intra-CCC) равно $s_2 = 3,50$. Среднеквадратичное отклонение разности $M_2 - M_1$ равно

$$s_d = \sqrt{s_1^2 + s_2^2} = 4,30.$$

F-отношение, меньшее 2, допускает использование t-критерия со значением статистики, равным

$$t = (M_2 - M_1)/s_d = 4,03.$$

Это t-значение при числе степеней свободы 58 делает соответствующее P-значение (т.е. вероятность такого же или большего CCC-отклонения) равным примерно 10^{-4}.



З а м е ч а н и е. В наших вычислениях мы предполагали, что inter-CCC различных кусков текста независимы. Нам представляется это разумной аппроксимацией. Intra-CCC могут иметь небольшую корреляцию. Например, выборочный коэффициент корреляции между первыми пятнадцатью и последними пятнадцатью intra-CCC части 1 равен только 0,156. Такая маленькая корреляция не может значительно изменить t-критерий.

Наши вычисления t-критерия по двум выборкам говорят о том, что две части написаны разными авторами (при довольно высоком уровне значимости). Результат нашего независимого от содержания исследования подтверждается аналогичным заключением с помощью лингвистического анализа (Бар-Селла [1]). Следует подчеркнуть, что результаты этих двух исследований основаны на различных свойствах текста и, таким образом, взаимно подтверждают друг друга.

Авторы благодарны за помощь, оказанную Д. Малютовым и И. Малютовым, а также за совет Зеева Бар-Селлы по выбору приложения.

СПИСОК ЛИТЕРАТУРЫ

1. *Бар-Селла З.* Литературный котлован: проект «Писатель Шолохов». М.: РГГУ, 2005.
2. *Benedetto D., Caglioti E., Loreto V.* Language trees and zipping. — Phys. Rev. Lett., 2002, v. 88, № 4, 28 January 2002, p. 048702.
3. *Bennett C. H., Gács P., Li M., Vitanyi P. M. B., Zurek W.* Information Distance. — IEEE Trans. Inform. Theory, 1998, v. 44, № 4, p. 1407–1423.
4. *Chomsky N.* Three models for the description of language. — IRE Trans. Inform. Theory, 1956, v. 2:3, p. 113–124.
5. *Cilibrasi R., Vitanyi P.* Clustering by compression. — IEEE Trans. Inform. Theory, 2005, v. IT-51:4, p. 1523–1545.
6. *Elliott W. Y., Valenza R. J.* And then there were none: winning the Shakespeare claimants. — Computers and the Humanities, 1996, v. 30, p. 191–245.
7. *Фоменко В. П., Фоменко Т. Г.* Авторский инвариант русских литературных текстов. Кто был автором «Тихого Дона»? — В кн.: *Фоменко А. Т.* Методы статистического анализа исторических текстов. Приложения к хронологии. Т. 2. Приложение 3. М.: Изд-во Крафт+Леан, 1999.
8. *Gelbukh A., Sidorov G.* Zipf and Heaps laws coefficients depend on language. — Lect. Notes Comput. Sci. 2001, B. 2004, S. 332–335.
9. *Колмогоров А. Н.* Три подхода к определению понятия «количества информации». — Проблемы передачи информации, 1965, т. 1, в. 1, р. 3–11.
10. *Кукушкина О. В., Поликарпов А. А., Хмелев Д. В.* Определение авторства текста с использованием буквенной и грамматической информации. — Проблемы передачи информации, 2001, т. 37, в. 2, с. 96–108.
11. *Li M., Chen X., Li X., Ma B., Vitanyi P.* The similarity metric. — IEEE Trans. Inform. Theory, 2004, v. 50, № 12, p. 3250–3264.
12. *Малютов М. Б.* Обзор методов и примеров атрибуции текстов. — Обозрение прикл. и промышл. матем., 2005, т. 12, в. 1, с. 41–77.
13. *Malyutov M. B., Wickramasinghe C. I., Li S.* Conditional complexity of compression for authorship attribution. — SFB 649 Discussion Paper № 57. Berlin: Humboldt University, 2007.
14. *Маслов В. П., Маслова Т. В.* О законе Ципфа и ранговых распределениях в лингвистике и семиотике. — Матем. заметки, 2006, т. 80, № 5, с. 718—732.
15. *Mendenhall T. A.* The characteristic curves of composition. — Science, 1887, v. 11, p. 237–249.
16. *Mendenhall T. A.* A mechanical solution to a literary problem. — Popular Science Monthly, 1901, v. 60, p. 97–105.
17. *Merhav N.* The MDL principle for piecewise stationary sources. — IEEE Trans. Inform. Theory, 1993, v. 39, № 6, p. 1962–1967.
18. *Rissanen J.* Universal coding, information, prediction and estimation. — IEEE Trans. Inform. Theory, 1984, v. 30, № 4, p. 629–636.
19. *Rosenfeld R.* A maximum entropy approach to adaptive statistical language modeling. — Computer, Speech and Language, 1996, v. 10, p. 187–228.
20. *Ryabko B., Astola Y.* Universal codes as a basis for time series testing. — Statist. Methodology, 2006, v. 3, p. 375–397.
21. *Shannon C.* Communication theory of secrecy systems. — Bell System Techn. J., 1949, v. 28, p. 656–715.

22. *Szpankowski W.* Average Case Analysis of Algorithms on Sequences. N. Y.: Wiley, 2001.
23. *Ziv J.* On classification and universal data compression. — IEEE Trans. Inform. Theory, 1988, v. 34, № 2, p. 278–286.
24. *Фитингоф Б. М.* Оптимальное кодирование при неизвестной и меняющейся статистике сообщений. — Проблемы передачи информации, 1966, т. 2, в. 2, с. 3–11.
25. *Фитингоф Б. М.* Сжатие дискретной информации. — Проблемы передачи информации, 1967, т. 3, в. 3, с. 28–36.
26. *Rocha J., Rosella F., Segura J.* The Universal Similarity Metric does not detect domain similarity. — arXiv:q-bio.QM/0603007 v1 6 Mar 2006.

Поступила в редакцию
24.III.2008