Использование информационной метрики в анализе текстового материала на примере корпуса текстов А. и Б. Стругацких

Д.Хмелёв

26 сентября 2004

1 Введение

На заре развития кибернетики широкую публику сильно занимали вопросы использования ЭВМ в творческой деятельности и в её анализе. Например, могут ли компьютеры сочинять мелодии, тексты, мыслить, говорить, понимать и критиковать тексты, а в конечном счёте могут ли они превзойти людей разумом. В 60-е годы многим казалось, что едва машины станут считать быстрее, они уже сравняются в возможностях с человеком. Однако, компьютеры оказались плохими учениками.

Неутомимое удвоение скорости вычислений в последние 30 лет ускорило компьютеры, но не помогло им добиться творческих успехов. Да, есть программы сочиняющие мелодии. Но они основаны на том, что имея перед собой несколько последних нот, компьютер обращается к известным мелодиям содержащим те же фрагменты и продолжает мелодию одним из возможных вариантов продолжения в этих фрагментах. Получаются мелодичные, но не цепляющие мелодии.

Сочинение текстов даётся компьютерам с ещё большей трудностью. По большому счёту, тексты получаются бессмысленными, хотя они и могут удовлетворять некоторым формальным критериям. Формальности настолько хорошо удаются программам, что результат вызывает изрядное веселье, например, в программах “Трепло” и “Стихоплюй” Михаила Гринчука: `http://shade.msu.ru/~taras/Grinchuk/`. Более современный

философ представлен проектом Яндекс/Весна: `http://www.yandex.ru/vesna.html`.

За полвека эволюции компьютеров, прояснилась неприглядная истина: компьютер умеет лишь то, что умеет сидящий за этим компьютером. Гуманитарии оказались в проигрышном положении по сравнению с представителями естественных наук, хотя я и читал как некоторые применяли упомянутый "Стихоплюй" чтобы менять направление размышлений. Любопытно, что другая часть гуманитариев пребывает в убеждении, что с помощью компьютеров можно сделать *всё*, и только их компьютерная безграмотность не позволяет написать программу, моментально извлекающую глубинную сущность из произведений, скажем, Н.В. Гоголя. И те и другие не правы. Компьютер можно применять для более интеллектуальных дел, чем сочинение бессмысленных стишков, но и ответ на *любой* вопрос тоже не получить не получится. В этой работе компьютер используется, чтобы определять *близость* между текстами.

Корректно определённая *количественная* мера близости текстов позволяет ответить на много интересных вопросов. Изменяя представление текстов можно изучать, какие факторы отличают одну группу текстов от другой. До сих пор *неизвестно*, можно ли *формально* отличить смешной текст от несмешного, увлекательный от занудного и т.п. Описанный далее метод представляет собой *инструмент*, с помощью которого можно пытаться решать такие вопросы. Причём инструмент этот пока *не откалиброван*. Таким образом, к представленным результатам следует относиться с осторожностью. Тем не менее, кажется достаточно интересным посмотреть, что же выдаст метод, получив на вход все произведения Стругацких, которые и так достаточно хорошо классифицированы. Можно ли автоматически получить представление о связи между текстами? Мне кажется, что да, и нам следует пробиться через несколько определений, чтобы понять суть метода.

2 О расстоянии и сжатии

Математически, близость текстов S и T характеризуется неотрицательным числом, которое называется *расстоянием*. Если это число велико, то тексты *далеки* друг от друга. Если же оно мало, то тексты *близки*. Расстояние, которое мы будем обозначать буквой d должно удовлетворять трём условиям:

(1) расстояние между любыми двумя текстами должно быть неотрицательным: $d(S,T) \geq 0$ на всех текстах S, T; и расстояние между текстами обращается в ноль, лишь когда они совпадают: $d(S,T) = 0$ лишь когда $S = T$;

(2) расстояние не меняется от перестановки текстов: $d(S,T) = d(T,S)$

(3) для любых трёх текстов S, T и N выполнено неравенство треугольника: $d(S,T) \leq d(S,N) + d(N,T)$.

Условие (3) можно уяснить на следующем примере: если лететь из Петербурга (S) в Тверь (T) напрямик, то получится быстрее, чем лететь сначала из Петербурга (S) в Нижний Новгород (N), а потом из Нижнего Новгорода (N) в Петербург (S). Другими словами, полёты без пересадок короче полётов с пересадками, если лететь одним и тем же самолётом.

Впервые расстояние между текстами пытался определить Н.А. Морозов [1] в 1915 году. Его наивная попытка состояла в том, что он считал количество словоупотреблений отдельных часто встречающихся слов, например союза “и” и предлога “в” в текстах разных авторов, и утверждал, что изучая количество словоупотреблений можно судить о том, принадлежит ли текст данному автору. Однако, если откладывать по оси “икс” относительное количество словоупотреблений “и”, а по оси “игрек” относительное количество словоупотреблений “в”, то точки, отвечающие разным текстом, окажутся вперемешку и близко друг к другу, так что отличить авторов друг от друга не получится. На последнее обстоятельство довольно скоро указал известный математик А.А. Марков [2]. Хотя Морозов и ткнул пальцем в небо, формально, он оказался провидцем. Некоторые современные методы классификации текстов пользуются аналогичной идеей, только считают количество словоупотреблений у *всех* слов в текстах, чтобы по комбинациям частот словоупотреблений авторов и различать (см. обзор [3]).

Мы воспользуемся подходом к определению расстояния, который пытается находить *содержательные* связи между текстами. Конечно, содержательные связи между текстами отражены в частотах словоупотреблений. Тем не менее автор придерживается той точки зрения, что тексты не следует редуцировать к словоупотреблениям. Другими словами, наилучший объект, характеризующий текст, это сам текст.

Информационное расстояние можно посчитать с помощью *программ сжатия*. Многие имеющие отношение к компьютерам сталкивались с программами, которые позволяют уменьшать объём места, занимаемого документами. Классический пакет сжатия — это Zip, но есть и более со-

верменные пакеты, которые часто жмут значительно лучше, например, RAR и 7ZIP. Сжатый файл — это набор инструкций для разжимающей программы, который позволяет без потерь восстановить исходный текст. Хотя у разных пакетов набор инструкций в файлах разный[1], размеры файлов получаются приблизительно одинаковыми[2]. Этот удивительный факт связан с тем, что современные компрессоры достигают почти максимальной степени сжатия текстов, оставляя минимальное количество информации. необходимой для того, чтобы буквально воспроизвести текст. Это минимальное количество называется *сложностью по Колмогорову* или *колмогоровской сложностью* [4] и обозначается $K(T)$ для текста T.

Чтобы определить сложность текста S относительно текста T нужно "подклеить" текст S к концу текста T и посмотреть, насколько хорошо сжимается эта добавка: $K(S|T) \approx K(TS) - K(T)$ см. также [5].

Относительная сложность, однако, не может служить метрикой, поскольку нарушаются условия (2) и (3). По условию (2) должно быть равенство $K(S,T) = K(T,S)$, но оно часто нарушается поскольку сложность $K(S|T)$ текста S маленькой длины относительно текста T большой длины уж конечно меньше $K(T|S)$. Если же брать арифметическое или геометрическое среднее $K(S|T)$ и $K(T|S)$, то получится нечто симметричное, но не удовлетворяющее условию (3) — неравенству треугольника. Тем не менее, относительную сложность можно успешно применять для классификации текстов по авторству, см. [5].

Подход к определению *настоящей* метрики предложен в работе [6]. Именно,

$$d(S,T) = \frac{\max(K(S|T), K(T|S)}{\max(K(S), K(T))}.$$

Авторам [6] удалось показать, что таким образом определённое расстояние почти удовлетворяет всем трём условиям. "Почти", потому что условия (1)–(3) выполняются с точностью до $1/\max(K(S), K(T))$, что становится несущественным при больших $K(S)$ и $K(T)$. В указанной работе [6] это расстояние было применено в задачах классификации видов по геному и языков по тексту декларации прав человека. Интересно, что

[1] поэтому не всегда одним архиватором можно разжать файл, сжатый другим; а иногда даже разные версии одной программы друг друга не понимают!

[2] особенно если пользоваться архиваторами, использующими наиболее современные методы — архиваторы типа Zip пользуются очень старым методом сжатия.

в классификации по геному вместо $K(S)$ можно использовать $N_k(S)$ — число неповторяющихся подстрок длины k, которые возможно перекрываются друг с другом в тексте, а k — небольшое число от 4 до 10. В работе [7] информация о частотах употребления пар последовательных букв текста успешно использована для классификации текстов по авторству. Метод статьи [7] использован в нашумевшем "Лингвоанализаторе": `http://www.rusf.ru/cgi-bin/fr.cgi`.

Некий сюрприз с метрикой $d(S,T)$ состоит в том, что она принимает значения между нулём и единицей. Вселенная текстов таким образом оказывается вмещена в круг диаметром единица и совсем удалённые друг от друга тексты оказываются на расстоянии порядка 1, которое, однако, никогда не достигается. Отличие от настоящего круга на плоскости состоит в том, что возможен миллион текстов, с попарными расстояниями близкими к единице. Этот неожиданный эффект называется "проклятием размерности": геометрия многомерных пространств чрезвычайно неинтутивна, и так же неинтуитивна метрика $d(S,T)$.

Мы применим расстояние $d(S,T)$ к классификации коллекции текстов А. и Б. Стругацких, и их подражателей.

3 Эксперимент на текстовом корпусе

Список произведений текстового корпуса приведён ниже. В нём перечислены основные произведения А. и Б. Стругацких, как совместные, так и по-одиночке под псевдонимами С. Ярославцев (А. Стругацкий) и С. Витицкий (Б. Стругацкий). Некоторые произведения присутствуют как в первоначально напечатанном, так и в каноническом вариантах. Кроме того, добавлены подражания разных авторов, в том числе и тексты неизвестного авторства:

Б Беспокойство (Улитка на склоне-1) (200k)
ВГВ Волны гасят ветер (292k)
ВНМ Второе нашествие марсиан (184k)
ГЛ Гадкие лебеди (392k)
ГО Град обреченный (771k)
ДР Далекая Радуга (234k)
ДСЛ Ярославцев С. Дьявол среди людей (201k)
ЖВМ Жук в муравейнике (365k)
ЖГП Жиды города Питера (87k)

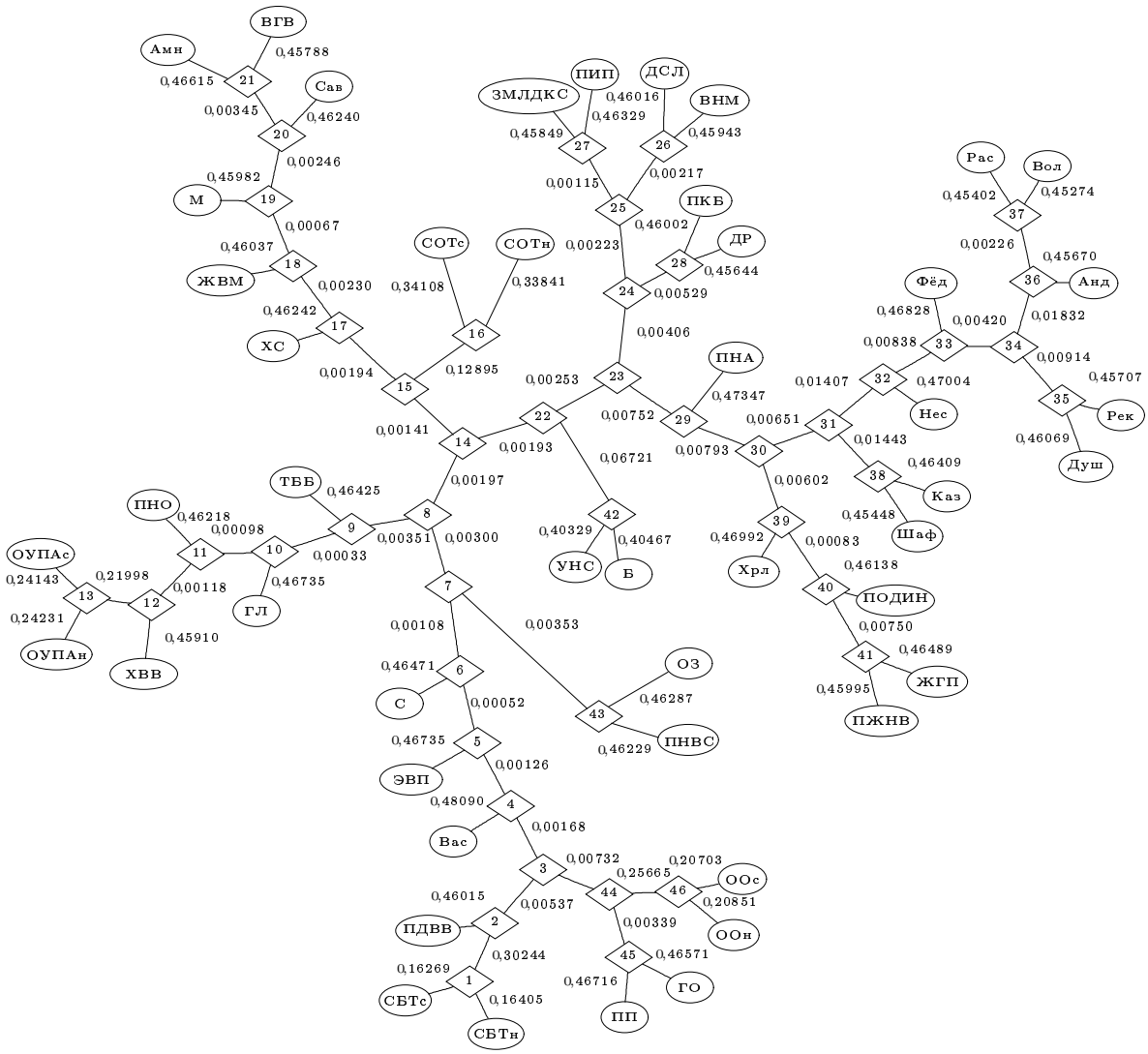

Рис. 1: Приближённое дерево расстояний между текстами

ЗМЛДКС За миллиард лет до конца света (220k)
М Малыш (288k)
ОЗ Отягощенные Злом (405k)
ООн Обитаемый остров (новый вариант) (658k)
ООс Обитаемый остров (старый вариант) (608k)
ОУПАн Отель “у Погибшего альпиниста” (новый) (360k)
ОУПАс Отель “у Погибшего альпиниста” (старый)(290k)
ПДВВ Полдень, XXII век. (480k)
ПЖНВ Ярославцев С. Подробности жизни Никиты Воронцова (71k)
ПИП Парень из преисподней (193k)
ПКБ Попытка к бегству (186k)
ПНА Путь на Амальтею (147k)
ПНВС Понедельник начинается в субботу (418k)
ПНО Пикник на обочине (318k)
ПОДИН Повесть о дружбе и недружбе (95k)
ПП Витицкий С. Поиск предназначения или 27-теорема этики (695k)
С Стажеры (385k)
СБТн Страна багровых туч (новый вариант)
СБТс Страна багровых туч (старый вариант)
СОТн Сказка о Тройке (новый вариант) (363k)
СОТс Сказка о Тройке (старый вариант) (219k)
ТББ Трудно быть богом (348k)
УНС Улитка на склоне (393k)
ХВВ Хищные вещи века (330k)
ХС Хромая судьба (сокр. журнальный вариант) (291k)
ЭВП Ярославцев С. Экспедиция в преисподнюю (437k)
Амн Амнуэль П. Лишь разумные свободны (296k)
Анд Андреева М. Трудно быть гадкой улиткой на склоне обочины в субботу (1,5k)
Вас Васильев В. Тень улитки (542k)
Вол Волин В. Фантасты о пришельцах (2,6k)
Душ Душенко К. Рукописи не горят (5k)
Каз Казаков В. Полет над гнездом лягушки (43k)
Нес Нестеренко Ю. Трудно быть багом, или Жук на обочине (24k)
Рас Раскин Ан. Трудно плыть боком (2,6k)
Рек (автор неизвестен) Неправильная рекурсия (7,9k)
Сав Савеличев В. Возлюби дальнего (273k)
Фёд Федоров И. соавт. Хищные Несси века (10k)

Хрл (автор неизвестен) Христолюди (74k)
Шаф Шафрановский Ю. Трудно быть странником (53k)

Были посчитаны попарные расстояния с помощью формулы для $d(S,T)$, причём для приближения колмогоровской сложности использовался алгоритм PPMD который разработал Д.Шкарин [8]. Предварительно тексты были преобразованы с помощью фильтра, оставлявшего лишь слова из букв кириллицы и пробелы между словами.

Расстояния удобно отображать с помощью дерева. Сумма длин расстояний на пути от одного листа до другого приблизительно равна настоящему расстоянию между листьями. Точного равенства не может быть, поскольку в таблице попарных расстояний между n текстами имеется $n(n-1)/2$ ненулевых разных чисел, а у деревьев такого вида с n листьями всего лишь $2n-3$ ребра. Поэтому с помощью специального алгоритма было построено дерево, у которого среднее квадратичное отклонение от истинных расстояний минимально.

4 Анализ результатов

4.1 Определение редакций одного текста

При первом взгляде на дерево, изображённое на Рис. 1 ясно, что разные редакции одного текста оказались близко друг к другу, причём расстояния дают возможность *количественно* определить степень редактирования, и сравнить, какие тексты редактировались больше.

- пара СБТс/СБТн, расстояние 0,32674=0,16269+0,16405
- пара ООс/ООн, расстояние 0,41554=0,20703+0,20851
- пара ОУПАн/ОУПАс, расстояние 0,48374=0,24143+0,24231
- пара СОТс/СОТн, расстояние 0,67949=0,34108+0,33841

Это находится в некотором соответствии с тем, как советская бюрократическая система заставляла авторов обращаться со своим текстом. Перечитайте “Сказку о Тройке”: это была совсем другая книга!

На этом фоне любопытен результат сравнения повести “Беспоойство” и романа “Улитка на склоне”:

- пара Б/УНС, расстояние 0,80796=0,40467+0,40329

Как видно, расстояние больше, чем между редакциями одного текста, но меньше, чем расстояние межу независимыми текстами, вроде ОЗ/С с расстояниям 0,93219=0,46287+0,00353+0,00108+0,46471

Таким образом, пара тексты Б/УНС *не являются* редакциями друг друга. Но говорить о том, что они совсем независимы тоже *нельзя*.

4.2 Группировка по тематике

С этого момента мы вступаем на зыбкую почву предположительных утверждений. Проблема с изображённым деревом заключается в том, что все расстояния между текстами, не являющимися редакциями друг друга за исключением пары Б/УНС, *приблизительно* одинаковы, и колеблются в пределах 0,92–0,96.

Таким образом, рёбра дерева с весом порядка 0,001 и меньше вполне могут быть проявлением случайности группировки. Следует понимать, что внутренние узлы (вершины) дерева служат лишь для передачи информации о расстояниях, и не означают ничего больше. Например, было бы ошибкой судить, что имеется сильная связь между парой СБТс/СБТн и книгой ПДВВ через узлы 1 и 2, поскольку ребро 1-2 имеет вес 0,30244. На самом деле, этот фрагмент дерева лишь означает, что обе редакции “Страны багровых туч” равномерно удалены от “мира Полдня”:

- d(ПДВВ, СБТс)=0,46015+0,30244+0,16269=0,92528
- d(ПДВВ, СБТн)=0,46015+0,30244+0,16405=0,92664.

Судя по небольшим числам на внутренних рёбрах, все существенно разные тексты оказались довольно равномерно далеки друг от друга. В этом нет ничего удивительного: было бы слишком скучно читать разные редакции одного текста, которые лежат близко друг к другу.

Между тем, очевидно, что определённый смысл в изображённом дереве *есть*, несмотря на то что какие-то рёбра могут стоять не на своих местах.

Во-первых, многочисленные подражатели оказались в основном оттеснены в отдельную гроздь, с корнем в вершине 31. Вклинивание подражателей Сав и Амн между узлом 19 и ВГВ может объяснятся значительной редакторской правкой ВГВ, да и текст Вас оказывается слишком

близко к группе текстов авантюрно-коммунистического характера ЭВП и С, которым легко подражать.

Очевидна также группировка текстов по тематике. Это гроздь вися-щая на узле 6, описывающая мир коммунистического будущего и его вариации на тему социалистического настоящего: ЭВП, С, ПДВВ, СБТс/СБТн, ООс/ООн, ГО и ПП. Характерно, что описание диктатур из "Обитаемого острова" и "Града обреченного" попало в коммунистическую группу.

О трудной судьбе интеллигента или сталкера в социалистической или капиталистической обстановке повествуют произведения грозди с корнем в узле 9: ТББ, ГЛ, ПНО, ХВВ, ОУПАс/ОУПАн. Естественно смотрятся поблизости ГЛ и ТББ с большим количеством застолий. Да и вообще, это группа произведений про жизнь "за границей".

Тройка произведений о Странниках и о Контакте отходит от узла 18: ЖВМ, М, ВГВ.

Борьба с бюрократией достойно отражена в тройке СОТс/СОТн и ХС, хотя эти произведения могут и оказаться просто случайно оказавшимися вместе, хотя лежащая неподалёку Улитка/Беспокойство, не представляется такой случайной.

"Космический" цикл отходит от узла 23, обрываясь на ПНА, и включает также ПКБ, ДР, ПИП, ЗМЛДКС, ВНМ, ДСЛ.

Искромётный юмор, особенно в описании всяких метафизических явлений несомненно привёл к сближению ОЗ/ПНВС: вспомните эпизоды описания приёмной Демиурга в ОЗ, или чемоданчик Агасфера Лукича, они явно отдают тёмными закутками ВНИИЧАВО из ПНВС.

Наконец, сгруппировались маргинальные ПЖНВ, ПОДИН и ЖГП. Ощущение, что группировка произошла в силу особенно большого количества диалогов в этих текстов

Любопытна позиция текста-подражания Хрл. По легенде, некто заметил стилистическое сходство между каким-то переводом и "Пикником на обочине" и подставил имена героев "Пикника". Но даже если сходство и было, то всё равно не такое уж и сильное, ибо текст Хрл оттеснён к группе остальных подражателей и нехарактерных текстов А. и Б. Стругацких.

Заметим, в заключение, что имеется пары Б/УНС и ОЗ/ПНВС произведений, которые не вписываются в указанные "серии" и подсоединяются к узлам 22 и 7, расположеным в центре графа. Поскольку информационная метрика d и применённый нами метод построения дерева исключительно формальны, некие формальные признаки выделяют эти

произведения на фоне всех остальных, и объясняют повышенный к ним интерес читателей, который и привёл к выходу настоящего сборника. Дальнейшее исследование, возможно, могло бы привести к некоему подобию "Изпитала" описанному в ХС. Однако, для разработки метода, который определяет практически значимый параметр "нкчт" (наиверо-ятнейшее количество читателей), потребуются усилия явно выходящие за рамки этой заметки.

5 Заключение

Подведём итоги. Во-первых, метод прекрасно обнаружил разные редакции одних и тех же текстов и даже показал насколько разными они являются. Во-вторых, построено логичное, хотя и малообоснованное, разбиение текстов Стругацких по "тематике". В третьих, оказалось что *стилистическая или тематическая близость являются эффектом второго, а то и третьего порядка*, но тем не менее явно обнаруживается информационной метрикой. По-видимому, незначительность этого эффекта и представляет основную трудность в применении компьютеров к гуманитарным исследованиям.

"Законы природы надо изучать", — утверждал Вечеровский, герой повести "За миллиард лет до конца света". Но изучению вычислительных методов в лингвистике, очевидно, препятствует Гомеостатическое Мироздание. В самом деле, математики почитают лингвистику в качестве разве что хобби, а на лингвистов, применяющих математические методы, смотрят как на маргиналов, да и маловато математически грамотных лингвистов. В результате и те и другие занимаются незаконным скрещиванием наук в свободное от основной работы время и получить финансирование на подобные исследования невозможно ни из естественно-научных, ни из гуманитарных фондов.

До тех пор, пока метод не будет опробован в тысячах ситуаций, и не станет ясно, что именно измеряет информационная метрика, *научные* выводы о близости будут страдать той же степенью неопределённости, что и настоящая статья. Тем не менее, ясно, что информационная метрика является интересным *инструментом* исследования, который наверняка будет применяться.

Список литературы

[1] Морозов Н.А. Лингвистические спектры: средство для отличения плагиатов от истинных произведений того или иного неизвестного автора. Стилеметрический этюд. // Известия отд. русского языка и словесности Имп.Акад.наук, Т.XX, кн.4, 1915. http://www.textology.ru/biblio.html

[2] Марков А.А. Об одном применении статистического метода. // Известия Имп.Акад.наук, серия VI, Т.X, N4, 1916, с.239. http://www.textology.ru/biblio.html

[3] Хмелёв Д.В. Классификация и разметка текстов с использованием методов сжатия данных. Краткое введение. http://www.compression.ru/download/articles/classif/intro.html

[4] Колмогоров А.Н. Три подхода к определению понятия количество информации. Пробл. Пер. Информ., 1(1):3-11, 1965 http://www.compression.ru/download/ti.html

[5] Кукушкина О. В., Поликарпов А., Хмелёв Д. Определение авторства текста с использованием буквенной и грамматической информации // Проблемы передачи информации. Т. 37, N 2. С. 96–108 2001.http://www.ma.utexas.edu/users/dima/PAPERS/khmelevr.html

[6] Ming Li, Xin Chen, Xin Li, Bin Ma, Paul Vitanyi The similarity metric. Выйдет в *IEEE Trans. Inform. Th.* http://www.compression.ru/download/articles/classif/articles.html

[7] Хмелёв Д. В. Распознавание автора текста с использованием цепей А.А. Маркова// Вестник МГУ, сер.9: филология. N2. С. 115-126, 2000. http://www.ma.utexas.edu/users/dima/PAPERS/khmelevr.html

[8] Д. Шкарин. Повышение эффективности алгоритма PPM. Проблемы передачи информации, Т. 37, N 3, С. 44-54, 2001. См. также: Библиотека сжатия PPMd. http://www.compression.ru/ds/