

УДК 621.391.1

О.В. КУКУШКИНА, А.А.ПОЛИКАРПОВ, Д.В. ХМЕЛЁВ

ОПРЕДЕЛЕНИЕ АВТОРСТВА ТЕКСТА С ИСПОЛЬЗОВАНИЕМ БУКВЕННОЙ И ГРАММАТИЧЕСКОЙ ИНФОРМАЦИИ

Статья опубликована в журнале “Проблемы передачи информации”, 2001, т.37, вып.2 (апрель-июнь), с.96–108. Translated in “Problems of Information Transmission”, pp. 172–184.

Поступила в редакцию 08.08.2000. После переработки 11.01.2001.

Содержание

1 Введение	2
2 Предварительная обработка	5
3 Метод и его перекрестная проверка	6
4 Описание результатов	8
5 Заключение	14
А Приложение	
Применение алгоритмов сжатия данных в задаче определения авторства	15

Метод, применяемый в данной статье для определения авторства текста, основывается на формальной математической модели встречаемости последовательности элементов текста как реализации цепи Маркова. В качестве элементов текста используются последовательности букв и последовательности грамматических классов слов. Оказывается, частоты употребления пар букв и пар грамматических классов в тексте на русском языке являются

достаточно устойчивой характеристикой автора и, видимо, их можно использовать, чтобы решать проблемы спорного авторства текста. Проводится сопоставление результатов, полученных при использовании различных вариантов методики в указанных единицах. Эксперимент проводится на 385 текстах 82 писателей.

В Приложении описано исследование Д.В. Хмелёва о возможности применения алгоритмов сжатия данных в задаче определения авторства.

1 Введение

В настоящей работе задача определения авторства ставится следующим образом. Пусть имеются достаточно длинные фрагменты прозаических произведений ряда авторов на русском или ином языке, использующем фонологическую (неиероглифическую) письменность¹. Про некоторый анонимный фрагмент известно, что он принадлежит одному из этих авторов, но какому — неизвестно. Требуется узнать, кому именно. На основе результатов экспериментальной проверки метода, предложенного Д.В. Хмелёвым в работе [1], утверждается, что с определенной, достаточно высокой степенью вероятности это возможно. Избранный метод базируется на учете статистики употребления пар элементов любой природы, идущих друг за другом в тексте (букв, морфем, словоформ и т.п.).

Традицию формального подхода к поиску методики определения авторства, по-видимому, следует отсчитывать от работы [2]. На эту работу откликнулся Марков [3], что свидетельствует о живом интересе создателя аппарата цепей Маркова к данной теме. Заметим также, что первое применение «испытаний, связанных в цепь», Марков описал в работе [4], посвященной анализу распределения гласных и согласных среди первых 20000 букв «Евгения Онегина».

Современное состояние методов определения авторства в нашей стране отражено в [5, гл. 5], а хороший обзор зарубежных авторов дан в работе [6]. Несмотря на огромное разнообразие описанных методов, ни один из них никогда не применялся к большому количеству текстов. Дело в том, что часто эти методы не поддаются автоматизации и требуют некоторого

¹Упоминание здесь текста, использующего именно фонологическую письменность, связано с тем, что иероглифическая система письма предоставляет меньше возможностей в отношении анализа парных встречаемостей элементов, так как фонологическая информация (различающая оболочки морфем и слов) в этом случае практически полностью скрыта за условной иероглифической оболочкой этих единиц.

человеческого вмешательства, что приводит к практической невозможности обработки большого количества текстов большого объема. В связи с этими проблемами встает вопрос общности каждого из методов: можно ли применять какой-либо из них вне ситуации, в которой они были разработаны?

Примечательным исключением до последнего времени являлась работа [7], в которой избранный там метод был применен к достаточно большому количеству текстов. В ней исследовалась доля служебных слов, которые используются автором, и обнаружено, что она устойчива для каждого из авторов на большом количестве текстов русских писателей XVIII–XX-го веков. Данную методику авторы работы [7] применили к проблеме определения плагиата.

Новый метод определения авторства текстов, написанных на естественном языке (независимо от того, на каком именно), впервые предложен Хмелёвым в работе [1].

Новый метод основывается на формальной математической модели последовательности букв (и любых других элементов) текста как реализации цепи Маркова. По тем произведениям автора, которые достоверно им созданы, вычисляется матрица переходных частот употребления пар элементов (букв, грамматических классов слов и т.п.). Она служит оценкой матрицы вероятности перехода из элемента в элемент. Матрица переходных частот строится для каждого автора. Для каждого автора оценивается вероятность того, что именно он написал анонимный текст (или фрагмент текста). Автором анонимного текста полагается тот, у которого вычисленная оценка вероятности больше (т.е. используется принцип максимального правдоподобия).

Такой метод, как показывает его первое использование [1] в приложении к разнообразному материалу, демонстрирует свою удивительную точность. Тем более удивительно, что хорошие результаты получились при исследовании переходов буквы в букву.

Марковские цепи разных порядков использовались в многочисленных работах 50-60-х годов XX века, описанных в книге [8], для оценки энтропии различных типов текстов. Однако ни в одной из этих работ не поднимался вопрос о применении марковских цепей в рамках задачи определения авторства. Более того, общепринятой была точка зрения, что на уровне букв и буквосочетаний характеристики *любого литературного текста* близки к средним по языку и, с практической точки зрения, неразличимы (см. [8, с. 191, примечание] и [9, 3]). Принцип максимального правдоподобия также никогда ранее не применялся в задаче определения авторства текстов в силу следующих причин: во-первых, до появления компьютеров

и большого количества текстов в электронном виде было затруднительно производить подсчеты на материале большого объема; во-вторых, определённый психологический барьер представляло отсутствие теоретического обоснования такого подхода, поскольку цепь Маркова первого порядка может служить лишь первым и очень грубым приближением естественно-языкового текста, на что многократно указывалось в различных работах по оценке энтропии текста с помощью цепей Маркова больших порядков [8, с.197]; и, наконец, в-третьих, в самой задаче определения авторства считался перспективным поиск устойчивых количественных характеристик грамматического характера, позволяющих различать писателей, причем, по-видимому, единственный серьезный успех в этом направлении для русского языка был получен лишь недавно (см. [7]).

В настоящей работе мы развиваем процедуру проверки метода [1], а также проверяем возможности его применения с использованием разных единиц анализа:

(а) пар букв в их естественных последовательностях в тексте — в словах (в той форме, в которой они употреблены в тексте) и пробелах между ними;

(б) пар букв в последовательностях букв в приведенных (словарных, лемматизованных или исходных) формах слов; например, предыдущее предложение в таком случае предстает в виде «пара буква в последовательность буква приведенный словарный лемматизованный или исходный форма слово»;

(в) пар наиболее обобщенных («неполных») грамматических классов слов, частей речи, в их последовательностях в предложениях текста — существительные, глаголы, прилагательные и т.п.: всего 14 традиционно выделяемых грамматических классов слов — частей речи и 4 других условных категории вроде «конец предложения», «сокращение» «неясный класс» и «знак тире»; категория «неясный класс» введена в связи с тем, что разбор по грамматическим классам проводился автоматически (и покрывал более 99% всех встреченных слов), но некоторые слова (например, с опечатками) не поддавались автоматической обработке, а потому их грамматический класс оставался неясным.

(г) пар менее обобщенных («полных») грамматических классов слов (а именно таких семантико-грамматических разрядов, как одушевленные существительные, неодушевленные существительные, прилагательные качественные, относительные, притяжательные и т.п.).

Была произведена перекрестная проверка метода на материале 385

текстов 82 авторов (результаты проверки приведены в табл. 1 и 2, а список авторов с указанием объема материала и числа текстов данного автора — в табл. 3).

Одним из показателей точности метода может служить процент правильно определенных произведений. На материале вариантов (а) и (б) получены наиболее точные результаты (73% и 62% точных определений соответственно). На материале варианта (в) получен 61% точных определений. На материале варианта (г) получены существенно более худшие результаты (4%).

В §2 описываются принципы и результаты предварительной обработки текстов. В §3 описана методика перекрестной проверки. Детальное описание результатов дано в §4 с представлением выводов в заключении §5. В Приложении, которое написано Д.В. Хмельёвым, изложен еще один подход к определению авторства с использованием алгоритмов сжатия данных.

2 Предварительная обработка

Исходный корпус текстов в результате предварительной обработки был представлен во всех ранее описанных вариантах (а)–(г).

Отличие варианта (а) от первоначального текста (и в этом одно из отличий данного исследования от [1]) состоит в том, что отброшены все слова, для которых не удалось автоматически определить грамматический класс (а следовательно, и найти словарную форму). Это было сделано, чтобы можно было сравнивать результаты с результатами варианта (б). При этом весь текст превращен в последовательность слов и пробелов между ними, вся пунктуация была отброшена, и, кроме того, были выкинуты все слова с заглавной буквы (включая слова, с которых начинаются предложения; такая уловка, как показано в работе [1], позволяет значительно увеличить точность определения авторства; по-видимому, это улучшение связано с тем, что отбрасываются имена литературных героев, как правило, не соотносящиеся со стилем автора литературного текста). В используемом алфавите буква «ё» склеивалась с буквой «е», в результате чего вместе с пробелом получилось 33 буквы. Каждая буква кодировалась своим номером: буква «а» соответствовала 1, ..., «я» соответствовала 32. Пробелу сопоставлялся код 0. Общее количество буквоупотреблений составило 96209964. Общее число разных пар букв, обнаруженных на анализируемом массиве текстов, составило 1011 (из потенциально возможных $33 \times 33 = 1089$). Очевидно, 1011 несколько превосходит количество разных пар

букв, действительно встречаемых в текстах на русском языке. Это является результатом определенного количества опечаток в электронных версиях книг и может приводить к погрешностям в вычислениях. Чтобы получить некоторую оценку этого шума, были отобраны все буквосочетания, которые вряд ли встречаются в русском языке, и их оказалось 121. Общее количество употреблений этих буквосочетаний равно 38495, что составляет около 0,04% от общего числа употреблений буквосочетаний, чем вполне можно пренебречь.

Преобразование исходных текстов для получения их в виде вариантов (б), (в) и (г) осуществлено на основе автоматического классификатора, разработанного О.В. Кукушкиной в Лаборатории общей и компьютерной лексикологии и лексикографии филологического факультета МГУ на основе грамматического словаря Зализняка [10] и академических грамматик [11, 12].

Вариант (в) базируется на информации, получаемой при использовании самого общего грамматического класса словоформы, т.е. информации о части речи (частицы, предлоги, междометия, соединительные и противительные союзы, прочие союзы, глаголы, местоимения, наречия, прилагательные, существительные, числительные, предикативы, компаративы, модальные наречия).

Вариант (г) базируется на информации, получаемой при учете лексико-грамматического разряда слов данной части речи (одушевленные существительные, неодушевленные существительные и т.п.).

Приведем некоторую статистику по текстам, отвечающим вариантам (б)–(г).

В варианте (б) общее количество букв составило 110704464. Число разных пар букв в исходных формах слов составило 1029 (из потенциально возможных $33 \times 33 = 1089$). Как видно, оно увеличилось в сравнении с вариантом (а). Такой эффект связан с переводом косвенных форм в исходные.

В вариантах (в) и (г) число элементов составило 20262449. При этом в варианте (в) количество разных пар элементов — 302 (из потенциально возможных $18 \times 18 = 324$). В варианте (г) число разных пар элементов составило 8124 (из потенциально возможных $112 \times 112 = 12544$).

3 Метод и его перекрестная проверка

Анализ текстового материала по каждому из указанных выше вариантов произведен на основе комплекса программ, разработанного Хмелёвым. При обработке каждого варианта отображения корпуса текстов мы покажем результаты перекрестной проверки метода [1]. Перекрестная проверка выполняется следующим образом.

Напомним, что элементы текста кодируются числами от 0 до 32 (в вариантах (а) и (б)), 17 или 111 (в вариантах (в) и (г) соответственно). Код 0 всегда соответствует разграничителю между крупными единицами: в вариантах (а) и (в) код 0 соответствует пробелу между словами, а в вариантах (в) и (г) код 0 соответствует разделителю между предложениями («концу предложения»).

Пусть у нас есть W писателей, у каждого из которых есть N_w текстов, где $w = 0, \dots, W-1$. В первую очередь подсчитывается Q_{ij}^{wn} — количество переходов из буквы i в букву j в тексте n ($n = 0, \dots, N_w - 1$) автора w ($w = 0, \dots, W - 1$). Чтобы найти предсказание автора текста $\hat{n}$ (известного автора $\hat{w}$) с использованием информации об авторстве всех остальных текстов всех авторов (включая автора $\hat{w}$), мы подсчитываем

$$Q_{ij}^k = \sum_{n=0}^{N_w-1} Q_{ij}^{kn}, \quad Q_i^k = \sum_j Q_{ij}^k$$

для авторов $k \neq \hat{w}$, а для автора $\hat{w}$ мы исключаем текст $\hat{n}$ из обучающей выборки

$$Q_{ij}^{\hat{w}} = \sum_{n \neq \hat{n}} Q_{ij}^{\hat{w}n}, \quad Q_i^{\hat{w}} = \sum_j Q_{ij}^{\hat{w}}.$$

Теперь мы вычисляем

$$\Lambda_k(\hat{w}, \hat{n}) = - \sum_{i: Q_i^k > 0} \sum_{j: Q_{ij}^k > 0} Q_{ij}^{\hat{w}\hat{n}} \ln \frac{Q_{ij}^k}{Q_i^k}$$

и

$$\Lambda_{\hat{w}}(\hat{w}, \hat{n}) = - \sum_{i: Q_i^{\hat{w}} > 0} \sum_{j: Q_{ij}^{\hat{w}} > 0} Q_{ij}^{\hat{w}\hat{n}} \ln \frac{Q_{ij}^{\hat{w}}}{Q_i^{\hat{w}}}.$$

Если отвлечься от вырожденных случаев $Q_{ij}^k = 0$ и $Q_i^k = 0$, то легко увидеть, что каждое $\Lambda_k(\hat{w}, \hat{n})$ есть минус логарифм вероятности реализации

текста $\hat{n}$ писателя $\hat{w}$ при условии, что он является реализацией марковской цепи с переходными интенсивностями $P_{ij}^k = Q_{ij}^k/Q_i^k$. Обоснование для отбрасывания вырожденных слагаемых дают результаты об оптимальной оценке максимума правдоподобия, приведенные в [13, с.224].

Также определим ранг $R_k(\hat{w}, \hat{n})$ как ранг $\Lambda_k(\hat{w}, \hat{n})$ среди $\{\Lambda_k(\hat{w}, \hat{n}), k = 0, \dots, W-1\}$, причем мы отсчитываем начальный ранг с нуля: $R_k(\hat{w}, \hat{n}) \in \{0, \dots, W-1\}$; наименьший ранг соответствует наименьшему числу. Если текст соотнесен правильному автору, то ранг $R_{\hat{w}}(\hat{w}, \hat{n}) = 0$.

Если текст соотнесен какому-либо другому автору, а правильный автор оказался на втором месте, то $R_{\hat{w}}(\hat{w}, \hat{n}) = 1$, и т.д.

В результате перекрестной проверки мы получаем набор рангов

$$\{R_{\hat{w}}(\hat{w}, \hat{n})\}_{\hat{w} \in \{0, \dots, W-1\}, \hat{n} \in \{0, \dots, N_{\hat{w}}-1\}}.$$

Точность методики определения автора характеризуется этим набором рангов. Доля точных угадываний совпадает с долей нулевых рангов. Результаты, близкие к угаданному, характеризуются долей небольших рангов. В качестве характеристики точности угадывания может также служить среднее рангов:

$$M = \frac{1}{\sum_{\hat{w}=0}^{W-1} N_{\hat{w}}} \sum_{\hat{w}=0}^{W-1} \sum_{\hat{n}=0}^{N_{\hat{w}}-1} R_{\hat{w}}(\hat{w}, \hat{n}). \quad (1)$$

Мы также приведем результаты перекрестной проверки по методу сравнения частот одиночных букв (в вариантах (а) и (б)) и одиночных грамматических классов (в вариантах (в) и (г)). Этот метод работает так же, как уже описанный, но дополнительно рассчитываются величины

$$Q^k = \sum_i Q_i^k \text{ и } Q^{\hat{w}} = \sum_i Q_i^{\hat{w}}$$

и вместо величин $\Lambda_k(\hat{w}, \hat{n})$ используются величины

$$\Gamma_k(\hat{w}, \hat{n}) = - \sum_{i: Q_i^k > 0} \left(\sum_j Q_{ij}^{\hat{w}\hat{n}} \right) \ln \frac{Q_i^k}{Q^k}$$

и

$$\Gamma_{\hat{w}}(\hat{w}, \hat{n}) = - \sum_{i: Q_i^{\hat{w}} > 0} \left(\sum_j Q_{ij}^{\hat{w}\hat{n}} \right) \ln \frac{Q_i^{\hat{w}}}{Q^{\hat{w}}}.$$

Таблица 1: Результаты перекрестной проверки правильности определения автора текста с использованием последовательности букв

Случай (а)			Случай (б)		
Словоформы текста			Приведенные формы		
R	Пары букв	Одиночные	R	Пары букв	Одиночные
0	282/385	27/385	0	240/385	12/385
1	21/385	55/385	1	29/385	40/385
2	9/385	25/385	2	17/385	19/385
3	5/385	24/385	3	9/385	16/385
4	5/385	17/385	4	6/385	16/385
≥ 5	63/385	237/385	≥ 5	84/385	282/385
M	3,38	12,69	M	4,77	17,88

Таблица 2: Результаты перекрестной проверки правильности определения автора текста с использованием последовательности грамматических классов

Случай (в)			Случай (г)		
Обобщенные грамматические классы			«Полные» грамматические классы		
R	Парные	Одиночные	R	Парные	Одиночные
0	235/385	128/385	0	15/385	6/385
1	31/385	43/385	1	21/385	12/385
2	16/385	29/385	2	9/385	6/385
3	8/385	15/385	3	12/385	6/385
4	11/385	17/385	4	19/385	8/385
≥ 5	84/385	153/385	≥ 5 ,	309/385	347/385
M	5,43	10,13	M	17,76	31,93

4 Описание результатов

Результаты проведенного исследования представлены в табл. 1 и 2. Столбец R отвечает рангам 0, ..., 4 и рангам, не меньшим 5. В строке с рангом 0 приведена доля правильно определенных произведений, в строке с рангом 1 приведена доля таких произведений, что истинный автор оказался на втором месте, и т.д., и наконец, в строке ≥ 5 приведена доля таких произведений, что истинный автор оказался на месте не лучше шестого.

Строка M содержит средний ранг, определенный по формуле (1).

Сразу же отметим, что частоты одиночных букв и одиночных грамматических классов дают низкий, но все-таки существенно неслучайный уровень правильного определения авторства текстов (за исключением одиночных «полных» грамматических классов; причины этого нуждаются в дополнительном изучении).

Все расчеты с парами элементов (букв или грамматических классов) дают лучший результат по сравнению с расчетами по одиночным буквам и классам.

Теперь рассмотрим, чем различаются результаты анализа с использованием информации о статистике встречаемости пар букв в естественных словоформах в тексте от результатов использования статистики букв в приведенных (словарных) формах слов (см. табл. 1). Сопоставление показывает, что определение автора оказывается более успешным в случае работы со статистикой естественных последовательностей букв в тексте (73% против 62% точных определений авторства, и средний ранг 3,38 против 4,77). Видимо, «уравнивание» ряда словоформ одного слова при лемматизации снимает часть полезной для определения автора информации, связанной с окончаниями слов, что и приводит к ухудшению результатов в варианте (б) по сравнению со вариантом (а).

Сопоставление результатов определения авторства текста на основе статистики употребления пар букв и пар обобщенных (неполных) грамматических классов словоформ (варианты (а) и (в)) показывает, что результаты в том и в другом случае довольно высокие: 73% и 61% точных определений авторства соответственно. По среднему рангу успешного определения авторства обобщенные грамматические классы дают ошибку больше чем на 2 ранга. Относительно более низкая эффективность пар таких единиц как обобщенные грамматические классы (в сравнении с парами букв на естественных словоформах букв) на одном и том же материале может быть связана с меньшим объемом статистики по этим классам (как уже

упоминалось, на одном и том же корпусе текстов имеется около 96 миллионов буквоупотреблений в варианте (а) против 20 миллионов употреблений грамматических классов в варианте (в)).

Вместе с тем, обращает на себя внимание тот факт, что использование в анализе одиночных грамматических классов в варианте (в) (см. табл. 2) оказывается существенно более эффективным, чем использование одиночных букв: 33% успешных определений авторства в варианте (в) против 7% в варианте (а). Средний ранг в варианте (в) почти на два с половиной меньше среднего ранга варианта (а). Этот результат, видимо, связан с тем, что одиночные обобщенные грамматические классы слов в отличие от одиночных букв текста несут в себе более специфическую информацию, более точно характеризующую устойчивые структурные характеристики текстов каждого из авторов.

Наименее эффективным оказалось использование «полных» грамматических классов словоформ текста (вариант (г)) как в случае их попарного учета, так и поодиночке. Причины этого нуждаются в дополнительном исследовании.

В табл. 3 приведены данные об авторах анализируемых текстов, количестве их текстов, которые подверглись анализу, разбросе объемов текстов этих авторов с указанием минимального, среднего (в скобках) и максимального объемов, эффективности определения авторства во всех четырех вариантах ((а)–(г)) с указанием минимального, среднего (в скобках) и максимального ранга контрольного текста на его истинном авторе.

В экспериментах использовался также материал переводных текстов (С. Лем, И. Хмелевская, Б. Райнов). Любопытно, что результаты распознавания этих текстов не хуже, чем текстов, написанных авторами, у которых родной язык русский (хотя в выборку попали переводы Лема сделанные разными переводчиками).

Таблица 3: Количество и объем использованных текстов по писателям

№ п/п	Писатель	Кол-во текстов	Объем текстов (в тыс. букв)	(а)	(б)	(в)	(г)
0	О. Авраменко	7	223,7(279,5)395,1	0(0,0)0	0(0,0)0	0(0,0)0	10(13,9)22
1	А. Больных	7	0,8(185,0)298,8	0(4,1)24	0(5,9)37	1(12,1)79	4(15,3)75
2	К. Булычев	16	3,3(129,5)458,9	0(6,8)59	0(6,8)53	0(9,0)65	3(16,3)69
3	А. Волков	9	5,2(186,8)610,5	0(20,4)50	0(26,3)51	0(12,3)57	5(23,1)65
4	Г. Глазов	6	184,5(263,7)326,1	0(0,0)0	0(0,0)0	0(0,0)0	9(13,2)19
5	О. Гриневский	2	96,1(127,4)158,6	0(0,0)0	0(0,0)0	0(0,0)0	0(0,5)1

См. продолжение на след. странице.

№ п/п	Писатель	Кол-во текстов	Объем текстов (в тыс. букв)	(а)	(б)	(в)	(г)
6	Н.В. Гоголь	4	97,7(213,3)334,0	0(1,0)4	0(1,5)6	0(8,0)32	14(22,8)42
7	Н. Гумилев	2	70,1(70,6)71,0	0(0,0)0	0(0,0)0	0(0,0)0	0(0,0)0
8	Ф.М. Достоевский	4	88,6(175,5)268,9	0(0,0)0	1(2,0)3	0(0,3)1	10(13,3)18
9	М. и С. Дяченко	6	23,3(325,1)553,2	0(0,0)0	0(0,2)1	0(0,0)0	7(11,5)17
10	С. Есенин	2	44,6(131,5)218,4	0(0,0)0	0(0,0)0	0(0,0)0	0(0,5)1
11	А. Етоев	6	2,7(57,9)114,8	0(1,0)4	0(4,3)19	0(2,2)13	2(3,0)5
12	И. Ефремов	2	256,5(396,5)536,5	0(0,0)0	0(0,0)0	0(0,0)0	1(2,0)3
13	А. Житинский	3	253,6(793,2)1207,6	0(0,3)1	0(0,0)0	0(9,0)26	18(26,3)42
14	А. Кабаков	5	69,0(225,5)418,4	0(0,0)0	0(0,2)1	0(2,6)11	15(19,2)26
15	С. Казменко	5	132,8(400,5)1148,3	0(1,2)5	0(0,8)4	0(0,2)1	16(21,8)28
16	В. Каплан	7	19,3(91,9)305,2	0(4,1)25	0(5,6)24	0(5,0)23	9(23,0)40
17	А. Кац	2	81,7(461,4)841,0	0(0,0)0	0(0,0)0	0(0,0)0	1(2,5)4
18	В. Климов	4	58,5(107,5)179,9	0(7,0)15	0(7,0)20	0(1,5)6	3(6,8)9
19	Е. Козловский	2	848,6(868,4)888,2	0(0,0)0	0(2,0)4	15(42,0)69	40(56,5)73
20	И. Крашевский	3	380,6(555,2)803,1	0(0,0)0	0(0,0)0	0(0,0)0	6(8,3)10
21	И. Кублицкая	2	170,2(226,2)282,3	0(0,0)0	0(0,0)0	0(0,0)0	1(2,0)3
22	Л. Кудрявцев	4	108,3(190,5)348,2	0(0,3)1	0(1,5)5	0(0,0)0	8(13,5)24
23	В. Кунин	4	296,3(407,9)610,3	0(0,0)0	0(2,5)7	0(3,5)5	10(15,5)23
24	А. Курков	7	17,5(121,0)276,9	0(3,1)10	0(11,6)28	0(1,9)3	4(11,9)19
25	А. Лазаревич	4	11,3(101,3)274,7	0(14,3)47	5(20,3)54	2(11,0)18	4(9,3)15
26	А. Лазарчук	6	141,4(434,2)786,9	0(0,0)0	0(0,8)2	0(0,0)0	19(25,5)34
27	Ю. Латынина	3	116,8(970,7)2511,8	0(3,3)10	0(13,0)36	0(0,0)0	4(15,3)23
28	С. Лем	8	11,6(238,6)535,2	0(0,9)5	0(1,1)8	0(5,1)27	11(26,5)44
29	Н. Леонов	3	273,1(282,7)295,7	0(0,0)0	0(0,0)0	0(0,0)0	3(4,0)5
30	С. Логинов	14	1,3(153,4)916,2	0(15,9)36	4(18,1)37	0(18,9)49	14(35,6)59
31	Е. Лукин	5	26,9(144,6)367,9	0(3,2)15	0(0,0)0	0(4,4)19	8(16,0)39
32	Л. и Е. Лукины	4	105,2(239,9)564,7	0(0,3)1	2(3,8)6	0(0,5)2	4(12,8)20
33	С. Лукьяненко	15	6,0(277,6)542,9	0(3,0)22	0(6,9)76	0(9,1)58	9(25,4)73
34	Н. Маркина	2	93,6(179,8)266,0	0(0,0)0	0(0,0)0	1(2,5)4	0(1,5)3
35	А. Мелихов	2	457,6(536,4)615,2	0(0,0)0	0(2,5)5	0(0,0)0	17(17,5)18
36	В. Михайлов	2	84,2(169,3)254,5	0(0,0)0	0(0,0)0	0(0,0)0	1(2,5)4
37	А. Молчанов	2	206,5(302,4)398,3	0(0,0)0	0(0,0)0	0(0,5)1	4(5,5)7
38	В. Набоков	6	102,0(310,6)599,8	0(2,0)11	0(0,8)3	0(3,0)15	5(12,7)18
39	М. Наумова	4	5,2(161,1)337,8	0(7,8)31	0(11,8)47	0(17,8)69	4(10,0)17
40	Ю. Нестеренко	2	71,1(212,0)352,8	0(1,0)2	1(3,5)6	0(0,0)0	1(3,5)6
41	Ю. Никитин	3	656,9(681,4)702,2	0(11,3)34	0(17,0)51	0(0,7)1	5(6,7)8

См. продолжение на след. странице.

№ п/п	Писатель	Кол-во текстов	Объем текстов (в тыс. букв)	(а)	(б)	(в)	(г)
42	С. Павлов	2	375,6(414,5)453,4	0(0,0)0	0(0,5)1	0(0,0)0	6(6,5)7
43	А.С. Пушкин	2	57,1(113,7)170,3	0(0,0)0	0(0,0)0	0(0,0)0	0(0,0)0
44	Б. Райнов	5	267,7(363,6)420,3	0(0,0)0	0(0,0)0	0(0,6)3	12(13,8)15
45	Л. Резник	2	79,6(97,8)115,9	0(0,0)0	0(0,0)0	0(0,0)0	1(1,0)1
46	Н. Рерих	4	84,5(305,6)608,7	0(5,5)22	0(3,5)14	0(2,3)9	0(2,5)9
47	Н. Романецкий	7	5,5(203,2)530,6	0(5,4)21	0(7,7)20	0(1,0)4	15(27,7)45
48	А. Ромашов	2	87,7(88,1)88,4	0(0,0)0	0(0,0)0	2(3,0)4	0(0,0)0
49	В. Рыбаков	7	9,7(119,5)366,1	0(9,0)24	0(9,0)21	0(16,4)36	15(25,6)41
50	М.Е. Салтыков-Щедрин	2	101,6(170,4)239,1	0(0,0)0	0(0,0)0	0(0,0)0	1(2,5)4
51	Р. Светлов	3	29,2(241,0)425,4	0(0,0)0	3(13,3)20	0(0,7)2	3(8,7)17
52	А. Свиридов	4	13,4(224,0)601,5	10(27,5)65	0(11,8)41	0(11,0)44	6(26,5)56
53	В. Сегаль	3	60,5(132,0)259,7	0(0,0)0	0(0,0)0	0(0,3)1	1(4,7)8
54	К. Серафимов	2	75,3(130,8)186,4	0(0,0)0	0(0,0)0	0(0,0)0	1(1,0)1
55	И. Сергиевская	2	50,7(79,8)108,9	0(0,0)0	0(1,0)2	0(0,0)0	0(0,5)1
56	К. Ситников	8	13,0(66,1)274,3	0(0,0)0	0(2,3)18	0(0,5)4	3(8,3)22
57	С. Снегов	3	385,8(411,1)438,4	0(0,0)0	0(0,0)0	0(0,0)0	5(5,0)5
58	С.М. Соловьев	2	159,9(1251,6)2343,3	0(0,0)0	0(0,0)0	0(0,0)0	0(2,5)5
59	А. Степанов	6	83,7(219,6)390,3	0(0,0)0	0(0,0)0	0(0,0)0	4(5,0)8
60	А. Столяров	2	137,2(241,9)346,7	0(5,0)10	9(11,0)13	0(7,5)15	1(2,5)4
61	А. и Б. Стругацкие	30	37,1(230,4)579,5	0(1,9)24	0(2,5)23	0(5,9)54	15(36,4)79
62	А. Стругацкий	2	51,9(101,6)151,3	0(0,0)0	0(0,0)0	0(0,5)1	1(1,0)1
63	Б. Стругацкий	2	260,7(279,6)298,4	0(0,0)0	0(0,0)0	0(0,0)0	7(8,0)9
64	Е. Тильман	3	307,8(390,0)464,7	0(0,0)0	0(0,0)0	0(0,0)0	11(11,3)12
65	А. Толстой	2	97,9(113,8)129,7	0(0,0)0	0(0,0)0	0(0,0)0	0(0,0)0
66	Л.Н. Толстой	2	199,9(712,5)1225,1	0(0,0)0	0(0,0)0	0(0,0)0	1(1,5)2
67	Д. Трускиновская	9	82,6(235,9)478,6	0(0,8)3	0(2,7)12	0(3,8)28	13(23,8)53
68	А. Тюрин	19	1,3(222,0)832,7	0(2,3)20	0(1,2)13	0(2,6)25	24(34,0)55
69	Е. Федоров	2	221,3(667,2)1113,1	0(0,0)0	0(0,0)0	0(0,0)0	1(1,5)2
70	Е. Хаецкая	3	204,1(309,0)414,3	1(10,3)22	12(31,3)42	54(57,0)62	28(39,3)54
71	Д. Хармс	3	13,9(104,1)185,5	0(0,0)0	0(0,0)0	0(12,0)29	5(16,7)30
72	В. Хлумов	4	183,3(242,9)395,5	0(3,8)15	0(11,3)38	6(15,3)38	26(34,0)46
73	И. Хмелевская	5	203,7(345,7)459,1	0(0,0)0	0(0,0)0	0(0,4)2	9(12,8)18
74	В. Черняк	3	201,6(373,7)501,0	0(0,0)0	0(2,7)8	0(11,7)35	2(11,0)25
75	А.П. Чехов	3	247,9(335,3)414,5	0(0,0)0	0(0,0)0	4(13,3)20	18(18,7)20
76	В. Шинкарев	3	56,2(78,9)100,1	0(11,7)29	6(13,3)22	4(23,7)61	5(5,7)6

См. продолжение на след. странице.

№ п/п	Писатель	Кол-во текстов	Объем текстов (в тыс. букв)	(а)	(б)	(в)	(г)
77	В. Шукшин	2	66,7(187,7)308,8	0(0,0)0	0(0,0)0	0(0,0)0	1(2,5)4
78	С. Щеглов	3	55,2(103,0)146,1	0(4,0)12	0(13,7)41	0(3,0)9	2(2,3)3
79	А. Щеголев	3	105,6(318,0)561,7	0(0,0)0	0(0,0)0	0(0,0)0	12(18,7)32
80	В. Югов	6	66,7(149,2)304,3	0(0,5)2	0(0,7)2	0(1,8)10	8(10,7)18
81	В. Ян	2	507,3(553,9)600,4	0(0,0)0	0(0,0)0	0(0,0)0	2(2,0)2

5 Заключение

Основным результатом проведенного исследования является то, что использование грамматической информации в решении задачи определения действительного автора текста является не только осмысленным, но и достаточно эффективным, а в некоторых отношениях сопоставимым с использованием информации о встречаемости пар букв в тексте, как это было показано ранее в исследовании [1].

Вместе с тем продолжает вызывать удивление, что использование такой, казалось бы, простой единицы, как пара подряд идущих в тексте букв, дает более точные результаты, чем использование таких языковых категорий, как одиночные грамматические классы слов и их пары. Вполне возможно, что в буквенных парных структурах в преобразованном и, конечно, в неполном виде отображаются полные структуры морфем словоформ текста — префиксальные, корневые, суффиксальные и флексивные. Тем самым, довольно большой объем словоизменительной и словообразовательной информации о структуре русских слов оказывается отображенным в статистике парной встречаемости букв, что и определяет довольно высокий уровень эффективности использования этой статистики для определения авторства текста.

Другими словами, подсчет частот употреблений пар букв позволяет в некотором виде учесть информацию о словаре, который используется автором, а также, косвенно, информацию о предпочитаемых им грамматических конструкциях. Несмотря на то, что различия в частотах употреблений конкретных пар букв, скорее всего, несущественны, поскольку сходятся к частотам, средним по языку, при увеличении объема текстов (такой эффект был замечен еще Марковым [3]), «правдоподобие», учитывающее «общий» эффект изменения употреблений пар букв позволяет всё же с высокой степенью точности определить истинного автора произведения, что и было показано ранее в работе [1], а в настоящем исследовании подтверждено с помощью перекрестной проверки.

Однако вполне возможно, что дальнейшие эксперименты с грамматическими классами слов позволят за счет более точного грамматического анализа достичь более успешного результата, чем это оказалось возможным в рамках данного исследования. Перспективность использования грамматической информации, возможно, в нашем исследовании показана также и тем, что использование информации об одиночных обобщенных грамматических классах оказалось заметно более эффективным, чем использование информации об одиночных буквах.

То, что полученные в ходе исследования результаты с использованием различных единиц (букв и обобщенных грамматических классов) не противоречат друг другу, позволяет предположить, что в будущих развитых методиках определения авторства текста будут использоваться различные отображения текста, полученные на основе этих единиц для взаимной перепроверки результатов.

А Приложение

Применение алгоритмов сжатия данных в задаче определения авторства

В настоящем Приложении показано, как использовать алгоритмы сжатия данных для определения авторства и приведены результаты проверки этого нового метода определения авторства на примере той же выборки данных, что использовалась в [1] и в основном тексте статьи.

В [1] исследовалась выборка текстов 82 писателей. У каждого писателя был случайно отобран контрольный текст, а остальные тексты использовались в качестве обучающей выборки. После этого определялось авторство контрольного текста, и правильный ответ был получен в 69 случаях. Чтобы получить некоторую оценку того, насколько этот результат хорош в терминах вероятности правильного определения авторства, поставим следующий мысленный эксперимент. Предположим, что у нас имеется некоторый автомат, который при заданных двух текстах выводит 0, если это тексты разных авторов, и 1, если это тексты одного автора, причем вероятности ошибок первого и второго родов составляют некоторое число $0 < p < 1$. Рассмотрим применение этого автомата в рамках задачи

выбора правильного автора из 82. Тогда этот автомат должен вывести ровно 81 ноль, что отвечает сравнению контрольного текста с обучающими текстами неправильных авторов, и одну единицу, отвечающую сравнению контрольного текста с обучающим текстом истинного автора. Естественно рассматривать все остальные выводы как ошибочные. Тогда, если предположить, что все попарные сравнения независимы, вероятность правильного вывода составляет $(1 - p)^{82}$. При уровне доверия $p = 0,05$ мы получаем $(1 - 0,05)^{82} \approx 0,015$, уровень $p = 0,01$ отвечает $(1 - 0,01)^{82} \approx 0,439$, а при $p = 0,005$ получаем $(1 - 0,005)^{82} \approx 0,663$. Заметим, что $69/82 \approx 0,84$ и если мы хотим, чтобы наш гипотетический автомат превзошел по точности метод [1], необходимо потребовать, например, чтобы $p = 0,001$, и лишь тогда окажется $(1 - 0,001)^{82} \approx 0,921$. Таким образом, уровень правильного определения 69 контрольных текстов при выборе из 82 писателей следует расценивать как чрезвычайно высокий. С определенными оговорками это рассуждение применимо и к результатам основного текста статьи.

Под текстом мы будем здесь понимать последовательность символов из некоторого алфавита $\mathcal{A}$. Длину текста B мы будем обозначать через $|B|$. Назовем *конкатенацией* текстов B и A последовательность S длины $|B| + |A|$, начало которой совпадает с B , а конец совпадает с A . При этом будем писать, что $S = A.B$.

Дадим теперь «идеальное» определение относительной сложности в духе определения колмогоровской сложности (см. [14, 15]): относительная сложность $K(A | B)$ текста A относительно текста B — это длина наименьшей программы в двоичном алфавите, которая переводит текст B в текст A . К сожалению, $K(A | B)$ невычислима, и неясно, как можно ее использовать на практике.

Некоторое грубое приближение к ней (впрочем, вполне достаточное, как мы увидим далее, для целей определения авторства) можно получить с помощью алгоритмов сжатия данных. Определим *относительную сложность* $C(A | B)$ текста A по отношению к тексту B следующим образом. Сожмем текст B в текст B' , и текст $S = B.A$ в текст S' . Теперь положим $C(A | B) = |S'| - |B'|$. Данное определение содержит неоднозначность, ибо не сказано, каким именно способом производится сжатие. В настоящем исследовании будет использоваться несколько способов сжатия, которые описаны ниже.

Нас будет интересовать применение функции $C(A | B)$ в задаче определения авторства. Предположим, что у нас имеются тексты n авторов. Отберем у каждого автора по контрольному тексту $U_1, \dots, U_n$. Все остальные

тексты у каждого автора объединим в один текст $T_1, \dots, T_n$.

Для каждого контрольного текста i авторство определяется следующим образом. Сначала определяется ранг R_i числа $C(U_i | T_i)$ в наборе чисел $\{C(U_i | T_1), \dots, C(U_i | T_n)\}$. Ранги принимают значения от 0 до $n-1$. Если ранг R_i равен 0, то авторство i -го контрольного текста определено верно. Аналогично [1] можно ввести много различных характеристик точности метода определения авторства. Например:

1. Простейшая характеристика — число нулевых рангов R_i ;
2. Более обобщенную характеристику дает *средний ранг*

$$M = \frac{1}{n} \sum_{i=1}^n R_i.$$

Проверка различных алгоритмов сжатия проведена на корпусе текстов, который использовался в [1] и в основной части настоящей статьи. Как уже было упомянуто, корпус состоит из 385 текстов 82 писателей. Общий объем текстов составляет около 128 миллионов букв. Тексты подверглись предварительной обработке. Во-первых, были склеены все слова, разделенные переносом. Далее были выкинуты все слова, начинавшиеся с прописной буквы. Оставшиеся слова помещены в том порядке, в каком они находились в исходном тексте с разделителем из символа перевода строки. У каждого из $n = 82$ писателей было отобрано по контрольному произведению U_i . Остальные тексты были слиты в обучающие тексты T_i , $i = 1, \dots, 82$. Объем каждого контрольного произведения составлял не менее 50–100 тысяч букв.

Рассмотрим теперь возможные алгоритмы сжатия данных без потерь. Следующие алгоритмы наиболее популярны в последнее время: кодирование Хаффмана, арифметическое кодирование, метод Барроуза-Виллера [16] и множество вариаций метода Лемпеля-Зива [17]. Некоторые алгоритмы специально ориентированы на сжатие текста: это алгоритмы RPM [18] (использующие марковскую модель небольшого порядка) и DMC (использующий динамически изменяемую марковскую модель) [19]. Каждый алгоритм имеет большое число модификаций и параметров (например, существует динамическое кодирование Хаффмана, варьируется объем используемого словаря и пр.). Кроме того, существует множество «смешанных» алгоритмов, где текст, сжатый, например, с помощью алгоритма RPM, дополнительно кодируется с помощью кода Хаффмана.

Все эти алгоритмы реализованы в многочисленных программах, которых в настоящий момент существует не менее 150. Каждая из них реализует,

вообще говоря, разные варианты алгоритмов сжатия данных. Дополнительное разнообразие возникает из-за того, что у многих программ имеется несколько версий, которые также имеют разные алгоритмы сжатия. Было отобрано лишь несколько из них, широко представляющих весь спектр программ сжатия данных. Отобранные программы показаны в табл. 4.

Таблица 4: Программы сжатия

Программа	Автор	Используемые алгоритмы
1. 7zip версия 2.11	Игорь Павлов	арифм. кодирование, LZ + арифм. кодирование, PPM
2. arj версия 2.60	RAO Inc.	LZSS + Хаффман
3. bsa версия 1.9.3	Сергей Бабичев	LZ
4. bzip2	Джулиан Сьюард	Барроуз-Виллер + Хаффман
5. compress	Sun Inc.	LZW
6. dmc	Гордон В. Кормак	DMC
7. gzip	Жан-Луи Гаили	Шеннон-Фано, Хаффман
8. ha версия 0.999c	Гарри Хирвола	Скользящее окно словаря + арифм. кодирование
9. huff1	Вильям Демас (LDS 1.1)	статический Хаффман
10. lzari	Харахико Окумура (LDS 1.1)	LZSS+арифм. кодирование
11. lzss	Харахико Окумура (LDS 1.1)	LZSS
12. ppm	Вильям Теахан	PPM
13. ppmd5 версия F	Дмитрий Шкарин	PPM
14. rarw версия 2.00	Евгений Рошаль	вариант LZ77 + Хаффман
15. rar версия 2.70	Евгений Рошаль	вариант LZ77 + Хаффман
16. rk версия 1.03α	Малькольм Тейлор	LZ, PPMZ

Большинство из этих программ с описаниями можно получить из Индекса архиваторов², который поддерживает Джефф Гилхрист³. Программа **compress** взята из операционной системы SunOS 5.6. Программа **dmc**

²<http://web.act.by.net/~act/act-index.html>

³Jeff Gilchrist, jeffg@cips.ca

доступна по ftp⁴. Заметим, что dmc имеет опцию максимально используемой оперативной памяти. В нашем случае использовалась опция в 100000000 байт. Пакет программ LDS 1.1 также доступен по ftp⁵. Программа ppm доступна с домашней страницы ее автора⁶.

Результаты применения этих программ представлены в табл. 5, где в последней строке приведены данные, которые получаются при применении цепей Маркова [1] на том же материале. Вычисления, выполненные при составлении табл. 5, проводились под несколькими операционными системами на разных платформах и заняли около трех недель непрерывного счета.

Таблица 5: Точность определения авторства текста с использованием алгоритмов сжатия данных

Программа	Ранг						
	0	1	2	3	4	≥ 5	M
7zip	39	9	3	2	3	26	6,43
arj	46	5	2	7	2	20	5,16
bsa	44	9	3	1	1	24	5,30
bzip2	38	5	5	1		33	13,68
compress	12	1	1	3	2	63	24,37
dmc	36	4	3	4	4	31	9,81
gzip	50	4	1	2	1	24	4,55
ha	47	8	1	3	3	20	5,60
huff1	10	11	4	4	2	51	15,37
lzari	17	5	4	2	6	48	14,99
lzss	14	3	1	1	3	60	20,05
ppm	22	14	2	1	3	40	10,39
ppmd5	46	6	6	2		22	5,96
rar	58	1	1	1		21	7,22
rarw	71	3		2	1	5	1,44
rk	52	9	3	1		17	4,20
Цепи Маркова (см. [1])	69	3	2	1		7	2,35

Из данных табл. 5 следует, что программы сжатия угадывают истинных

⁴<http://plg.uwaterloo.ca/~ftp/dmc/>

⁵ftp://garbo.uwasa.fi/pc/programming/lds_11.zip

⁶<http://www.cs.waikato.ac.nz/~wjt/>

писателей весьма часто. Поэтому их использование, несомненно, имеет смысл. Заметим, что результаты применения программы **rarw** даже превосходят результаты, полученные ранее при использовании цепей Маркова [1]. Хотя такое превосходство можно отнести за счет определенной статистической погрешности, этот результат является лучшим на нынешний день.

По-видимому, такие прекрасные результаты связаны с тем, что программы сжатия действительно лучше адаптируются к контрольному тексту, переработав к тому времени обучающий текст истинного автора, чем какого-либо другого. Недостатком этого метода по сравнению с методом цепей Маркова [1] является то, что сами алгоритмы сжатия менее «прозрачны», а для коммерческих программ и вовсе недоступны для изучения. Тем не менее, многие среди представленных в табл. 4 программ, например, программы **gzip** и **bzip** имеют открытый исходный код и хорошо документированные открытые алгоритмы, и дальнейшее их изучение может открыть причины эффективности ложностного подхода к определению авторства.

Несомненное достоинство представленного здесь метода определения авторства состоит в том, что он доступен буквально на каждом компьютере и не требует никаких специальных программ при выборе из небольшого числа авторов, поскольку большинство из описанных архиваторов широко распространены, а некоторые (вроде **gzip** или **rar**) имеют реализации на всех платформах и во всех операционных системах.

СПИСОК ЛИТЕРАТУРЫ

- [1] Хмелёв Д.В. Распознавание автора текста с использованием цепей А.А. Маркова//Вестн. МГУ. Сер. 9, Филология. 2000. №2. С.115–126.
- [2] Морозов Н.А. Лингвистические спектры: средство для отличения плагиатов от истинных произведений того или иного известного автора. Стилеметрический этюд// Известия отд. русского языка и словесности Имп.акад.наук. 1915. Т.20, Кн.4.
- [3] Марков А.А. Об одном применении статистического метода//Изв.Имп.акад.наук, Сер. 6. 1916. №4, С.239–242.

- [4] Марков А.А., Пример статистического исследования над текстом “Евгения Онегина”, иллюстрирующий связь испытаний в цепь//Изв.Имп.акад.наук, Сер. 6. 1913. №3, С.153–162.
- [5] От Нестора до Фонвизина. Новые методы определения авторства. М.: Издат. группа “Прогресс”, 1994.
- [6] Holmes D.I. The Evolution of Stylometry in Humanities Scholarship//Literary and Linguistic Computing. 1998. V. 13. №3. P.111–117.
- [7] Фоменко В.П., Фоменко Т.Г. Авторский инвариант русских литературных текстов//Методы количественного анализа текстов нарративных источников. М.: Ин-т истории СССР, 1983. С.86–109.
- [8] Яглом А.М., Яглом И.М. Вероятность и информация. М.: Наука, 1960.
- [9] Добрушин Р.Л. Математические методы в лингвистике//Математическое просвещение. 1959. Вып.6.
- [10] Зализняк А.А. Грамматический словарь русского языка. М.: Рус. язык, 1977.
- [11] Грамматика современного русского литературного языка. М.: Наука. 1970.
- [12] Русская грамматика. Т.1,2. М.: Наука. 1980.
- [13] Ивченко Г.И., Медведев Ю.И. Математическая статистика. М.: Высш. шк., 1992.
- [14] Li M., Vitányi P. An Introduction to Kolmogorov Complexity and Its Applications. New York: Springer, 1997.
- [15] Колмогоров А.Н. Три подхода к определению понятия «количество информации»//Пробл. передачи информ. 1965. Т.1. №1, С.3–11.
- [16] Burrows M. Wheeler D.J. A block-sorting lossless data compression algorithm//Digital SRC Research Report 124. 1994. <ftp://ftp.digital.com/pub/DEC/SRC/research-reports/SRC-124.ps.gz>

- [17] Lempel A., Ziv J. On the Complexity of Finite Sequences//IEEE Trans. on Inform. Theory. 1976. V.22. №1. P.75–81.
- [18] Cleary J.G., Witten I.H. Data Compression Using Adaptive Coding and Partial String Matching//IEEE Trans. on Commun. 1984. V.32. №4. P.396–402.
- [19] Cormack G.V., Horspool R.N . Data Compression Using Dynamic Markov Modelling//Computer J. 1987. V.30. №6. P.541–550.