О роли устойчивых автокорреляций в текстах

д.т.н. Д.В. Ландэ, д.ф-м.н.А.А. Снарский
НТУУ «КПИ»

Задача нахождения критерия, позволяющего отличать текст от произвольного набора слов, сама по себе достаточно актуальна, например, в плане разработки средств индексирования интернет-контента [1] или разделения сигналов и шума в каналах связи [2]. Наиболее надежным критерием такого рода в настоящее время считается закон Ципфа [3]. Во всяком случае, случайные наборы слов этому закону не удовлетворяют.

Наиболее естественным подходом к решению поставленной выше задачи является, по-видимому, изучение автокорреляций в последовательностях слов, образующих текст. В предлагаемой работе обсуждается один из возможных критериев, основанный на определении степени сжатия данных. В качестве механизма изучения подобных автокорреляций может быть предложено сжатие данных, применяемое в различных системах сжатия файлов, в частности, в архиваторе ***gzip*** [4]. Действительно, в основе применяемого в нем метода сжатия лежит кодирование повторяющихся шаблонов [5], что, безусловно, в случае сжатия текстов зависит, и от порядка слов: чем больше в наборе данных устойчивых шаблонов, тем эффективнее сжатие.

Можно предположить, что если изменить порядок расположения слов в связном тексте, разрушив тем самым присутствовавшие там изначально шаблоны, уровень сжатия полученного результата уменьшится. При полном «перемешивании» слов в изначально связном тексте можно было бы ожидать установления некоторого значения уровня сжатия, которое уже слабо зависит от дальнейших перестановок слов.

Таким образом, изучая степень сжатия данных для различных состояний исходного документа, полученных путем последовательных перестановок слов, предположительно можно построить количественную меру степени «связанности» документа, которую естественно было бы считать одним из фундаментальных свойств текста.

Для проверки этой гипотезы авторами были проведены исследования ряда известных литературных произведений. Полученные в рамках работы результаты, хотя и требуют дальнейшего изучения, однако свидетельствуют в пользу принципиальной применимости метода.

В процессе исследований использовалась процедура перемешивания слов. Для определения уровня перемешивания были введены специальные правила. Пусть N – количество слов в тексте, следующих одно за другим. Под одной «атомарной» операцией

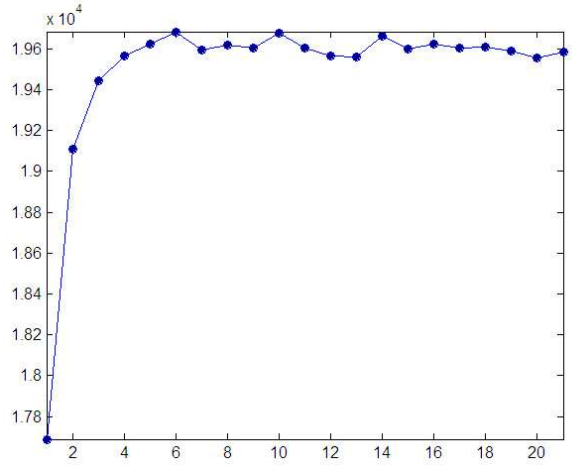
перестановки слов понимается следующая процедура: с помощью генератора псевдослучайных чисел определяются два числа n и m в интервале значений $[1, N]$. Далее, слова с порядковыми номерами n и m меняются местами.

В проводимых исследованиях выполнялось последовательно $\left[\frac{kN}{10}\right]$ «атомарных» перестановок слов ($k = 0,...,20; k = 0$ соответствует исходному состоянию). Таким образом, фиксируется 21 состояние исходного документа. Документ в каждом из полученных состояний подвергался сжатию с помощью алгоритма Лемпеля-Зива [4, 5], после чего определялся его объем, соответствующий данному состоянию. В результате были получены зависимости объемов сжатых текстов от степени перемешивания слов. Типичный вид таких зависимостей приведен на рис. 1. Заметим, что данная операция применялась к текстам с различными длинами.

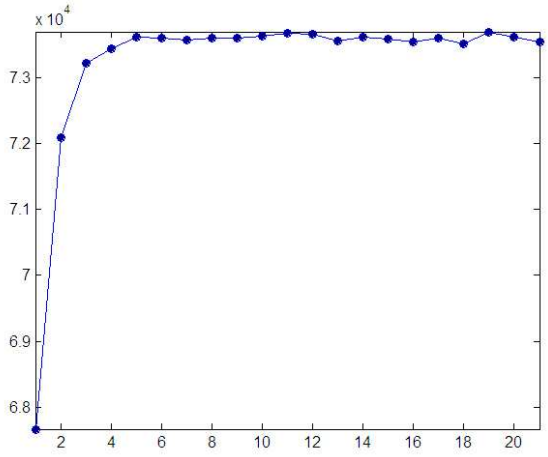
Мы видим, что в пределах $k < 6$ кривая демонстрирует монотонный рост, а затем выходит на насыщение с незначительными случайными колебаниями около некоторого среднего значения. Их амплитуда составляет около 10% от полной разности между максимальным и минимальным уровнями сжатия. Учитывая, что сжатый документ в среднем составляет порядка 30% от исходного, полученный эффект следует считать явно выраженным.

На рис. 2. приведены значения χ (отношения среднего установившегося значения объема сжатого перемешанного текста к объему сжатого исходного текста) в зависимости от длины сжимаемого фрагмента для различных текстов – произведений различных авторов. Как видно, значения χ возрастают при увеличении объемов текстов, и характер динамики этих значений, по-видимому, может свидетельствовать о стилях авторов.

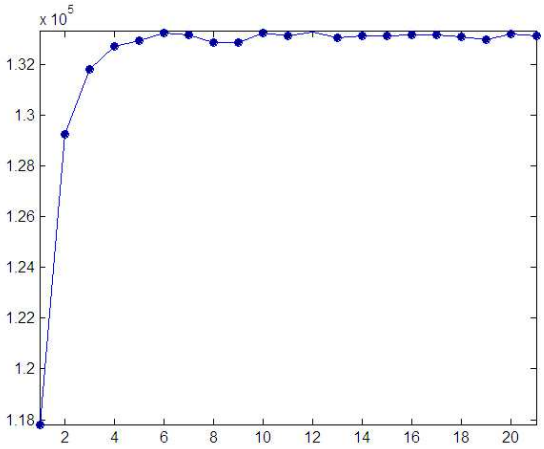
Были проведены исследования поведения параметров χ для текстов сообщений электронных СМИ. Как экспериментальная база был использован корпус сообщений электронных СМИ объемом 5000 документов. Средняя длина документов корпуса значительно меньше рассмотренных выше литературных произведений и составила всего 3951 символ. На рис. 3 приведен график зависимости χ (ось ординат) от номеров документов, ранжированных по данному параметру (ось абсцисс). Оказалось, что критерию $\chi > 1$ удовлетворяет 4909 документов, что составляет 98.18%. Примечательно, что аномально низкими χ обладают документы с наименьшей длиной (средняя длина 91 документа с $\chi \leq 1$ составляет всего 834 символа).



a)



b)



c)

Рис.1. Объемы архивов (ось ординат) в зависимости от коэффициента перемешивания: a) – Richard Bach. Jonathan Livingston Seagull (53149 символов); b) – William Shakespeare. Hamlet (207402 символов) ; c) – Ernest Hemingway. Green hills of Africa (363513 символов)

Авторами также был выполнен эксперимент по генерации искусственного текста. Для этого генерировались «искусственные» слова, содержащие от 1 до 12 букв. Из этих слов формировались тексты длиной около 10000 символов. При этом в формируемых текстах обеспечивались повторения отдельных слов с частотами, которые позволили смоделировать соотношения, установленные Ципфом. На рис. 4 приведен результат

сравнения массивов значений параметра χ для реальных и искусственных текстов. По оси абсцисс отложены номера исследуемых документов.

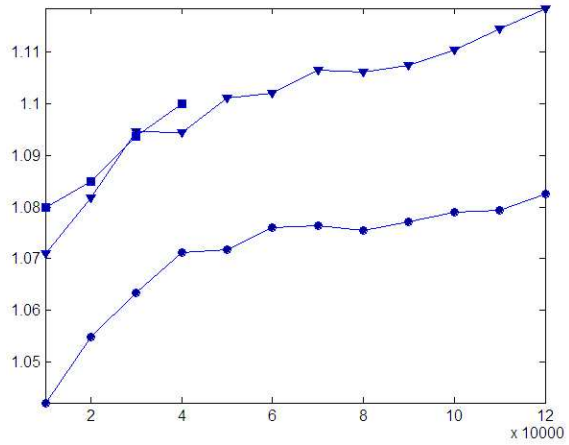


Рис. 2. Значения χ в зависимости от длины фрагмента:
■ – Richard Bach. Jonathan Livingston Seagull; ▼ – William Shakespeare. Hamlet ;
● – Ernest Hemingway. Green hills of Africa

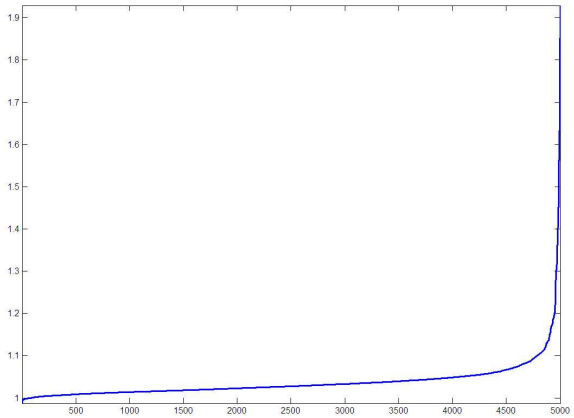


Рис.3. Ранговое распределение χ

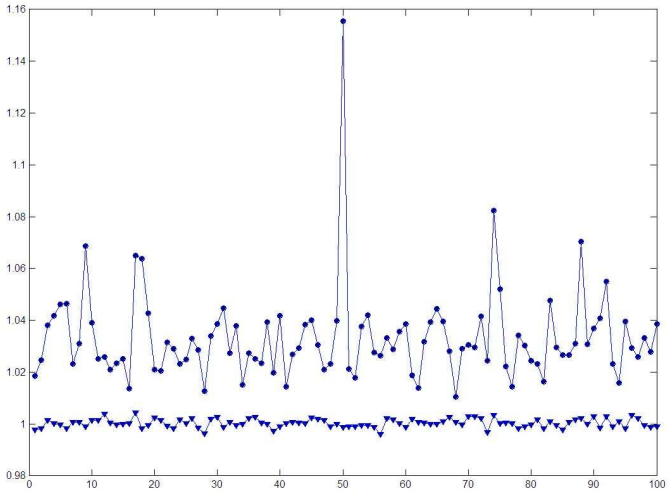


Рис. 4. Пераметр χ для реальных (•) и искусственных (▼) текстов

В ходе исследований удалось выявить формализуемое различие между случайными наборами слов и реальными текстами. Полученная зависимость объема сжатого текста от

степени перемешивания оказалась одним из критериев определения, является ли массив слов текстом либо просто набором слов, пусть даже формально удовлетворяющим закону Ципфа.

Литература

1. R. Baeza-Yates, P. Boldi, J.M. Gómez-Hidalgo. Adversarial Information Retrieval in the Web. UPGRADE (European Journal for the Informatics Professional), Monograph: Next Generation Web Search, Vol. VIII, issue №. 1, February 2007, pp. 33-40.
2. C. E. Shannon, W. Weaver. The Mathematical Theory of Communication. Univ of Illinois Press, 1949.
3. G.K. Zipf. Human Behavior and the Principle of Least Effort, Addison-Wesley, Cambridge: Univer. Press, 1949.
4. RFC 1952. «GZIP file format specification version 4.3.» - 1996. - 12 с.
5. J. Ziv, A. Lempel. A Universal Algorithm for Sequential Data Compression, IEEE Transactions on Information Theory, Vol. 23, №. 3, pp. 337-343.