МИБ - 2007

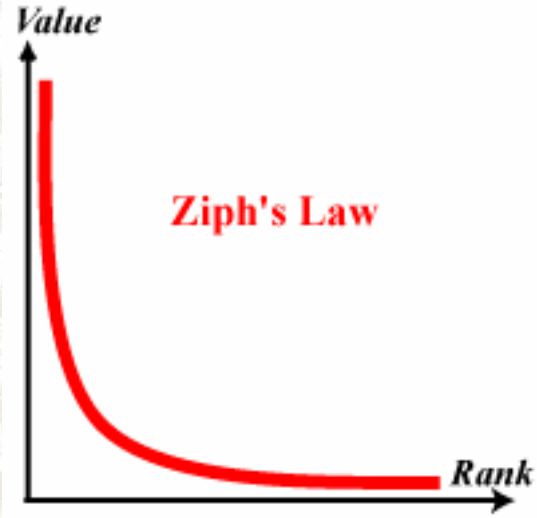
О РОЛИ УСТОЙЧИВЫХ АВТОКОРРЕЛЯЦИЙ В ТЕКСТАХ

Д.В. Ландэ, А.А. Снарский

ИЦ «ЭЛВИСТИ», НТУУ «КПИ»

Критерий, позволяющий отличать текст от произвольного набора слов

Закон Ципфа
Если к какому-либо достаточно большому тексту составить список всех встретившихся в нем слов, а затем отранжировать эти слова в порядке убывания частоты их встречаемости в тексте, то для любого слова произведение его ранга и частоты встречаемости будет величиной постоянной: f * r = c.

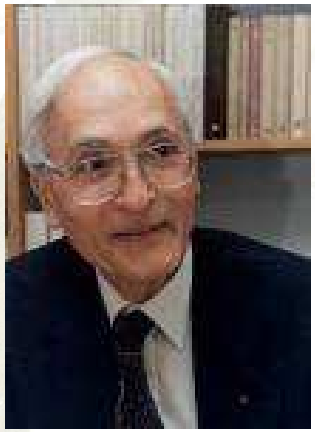


Джордж Зипф

Авторами обсуждается еще 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.

Исходное предположение

В основе применяемого при сжатии метода сжатия (алгоритма Лемпеля-Зива) лежит кодирование повторяющихся шаблонов, что, безусловно, в случае сжатия текстов зависит, и от порядка слов: чем больше в наборе данных устойчивых шаблонов, тем эффективнее сжатие.



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.

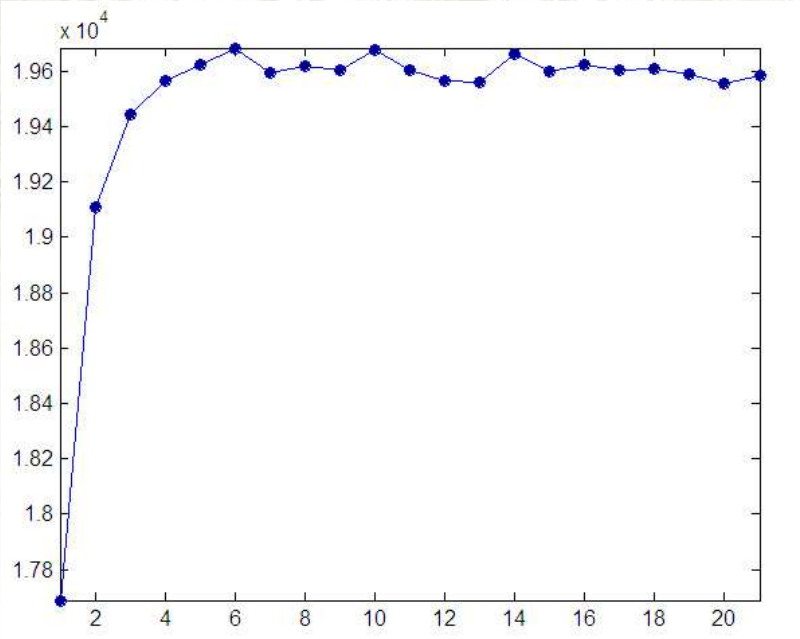
Перемешивание слов в текстах

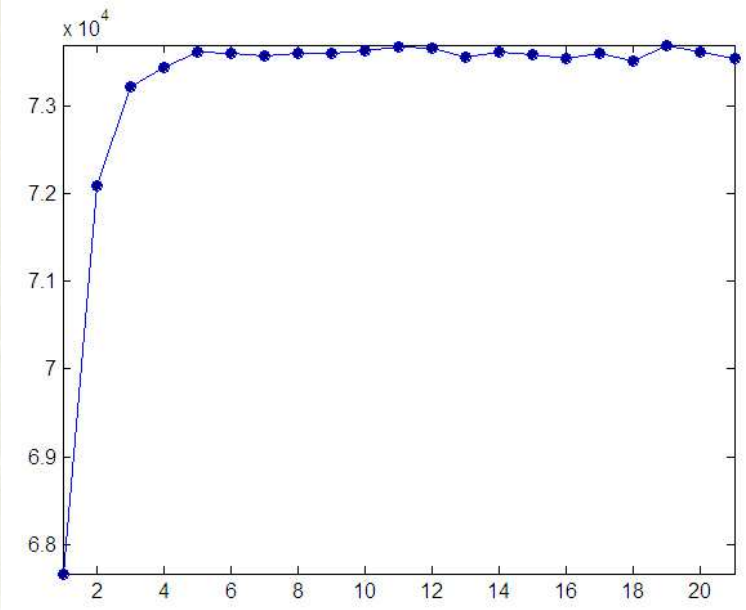
Можно предположить, что если изменить порядок расположения слов в связном тексте, разрушив тем самым присутствовавшие там изначально шаблоны, уровень сжатия полученного результата уменьшится. При полном «перемешивании» слов в изначально связном тексте можно было бы ожидать установления некоторого значения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

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.

Объемы архивов в зависимости от уровня перемешивания

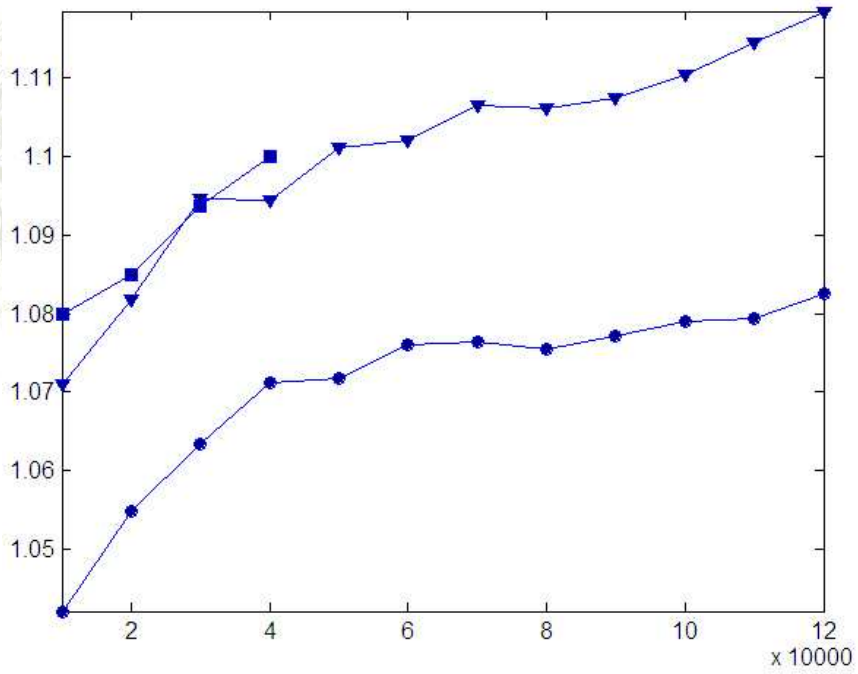
В проводимых исследованиях выполнялось последовательно [kN/10] «атомарных» перестановок слов (k=0,20). Таким образом, фиксируется 21 состояние исходного документа. Документ в каждом из полученных состояний подвергался сжатию с помощью алгоритма Лемпеля-Зива, после чего определялся его объем, соответствующий данному состоянию. В результате были получены зависимости объемов сжатых текстов от степени перемешивания слов.





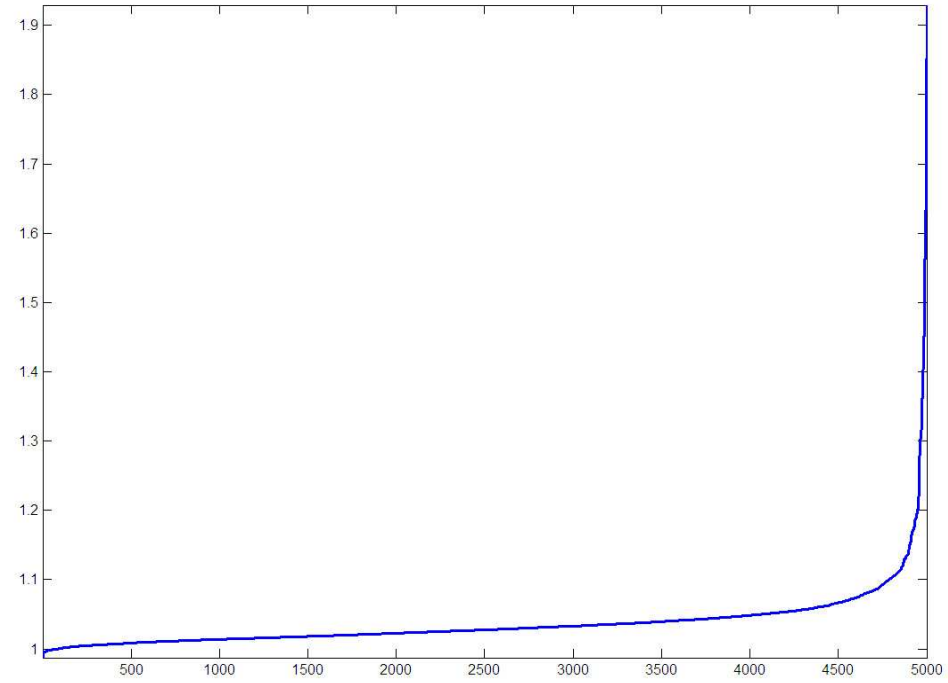
Параметр χ для литературных произведений

На рис. 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.



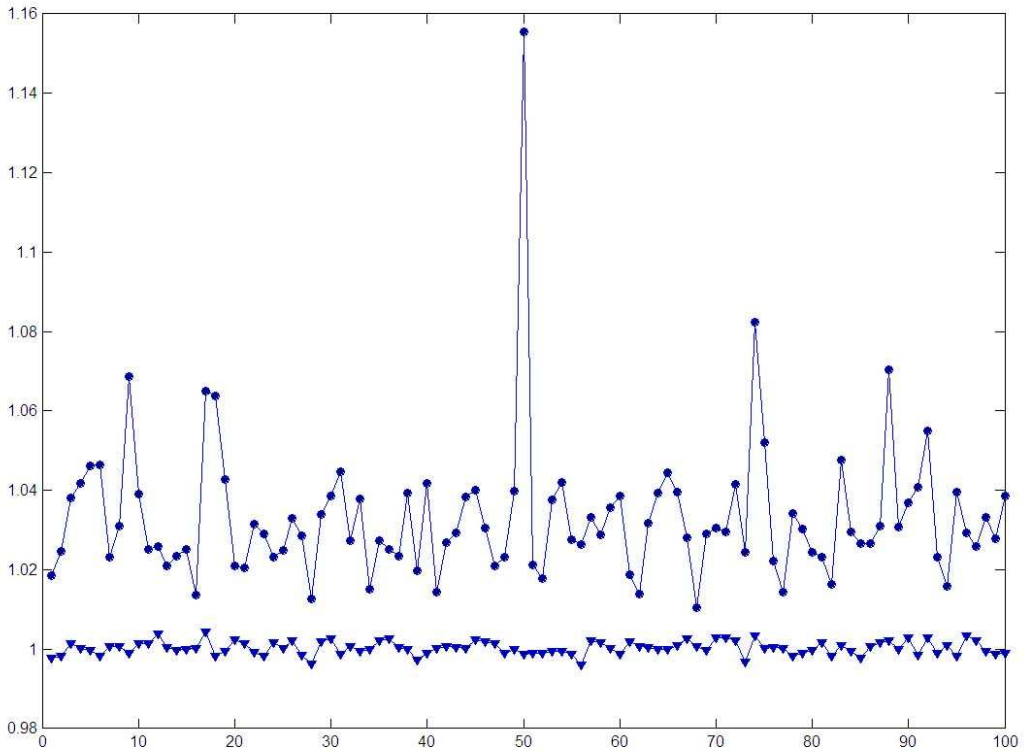
Параметр χ для массивов электронных СМИ

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. 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).



Параметр χ для реальных и искуственных текстов

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.



Заключение

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.

Спасибо за внимание!