ОСОБЕННОСТИ РАСПРЕДЕЛЕНИЯ УНИКАЛЬНЫХ СЛОВ В ТЕКСТОВЫХ МАССИВАХ

Ландэ *Дмитрий Владимирович*

ИПРИ НАН Украины, НТУУ «Киевский политехнический институт»
Киев, Украина
dwl@visti.net

Снарский *Андрей Александрович*

НТУУ «Киевский политехнический институт»
Киев, Украина
asnarskii@gmail.com

В докладе описывается подход к автоматическому определению жанра текстовых документов, основывающийся на анализе средней частоты встречаемости новых слов в заданном заранее «окне наблюдений».

Если последовательно каждому уникальному слову из текстового массива, начиная с первого, приписывать номер, то можно получить зависимость между позицией слова в тексте и этим номером. На рис. 1 приведен график этой зависимости для потока текстов из Интернета. Близкие по виду графики были получены и при анализе литературных произведений.

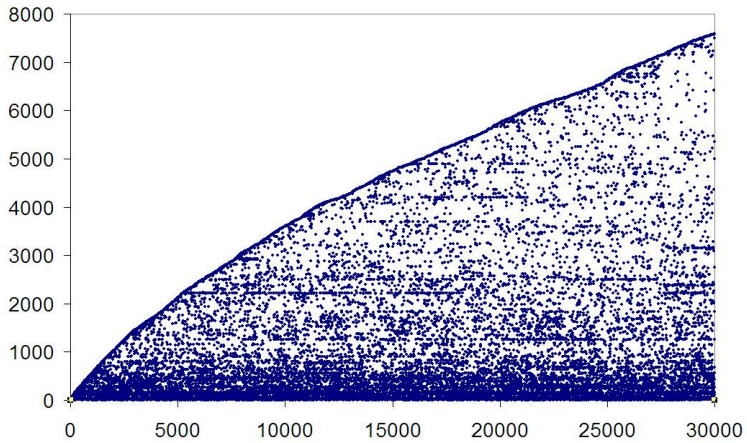

Рис. 1. График зависимости «номер слова в тексте – номер уникального слова»

Более высокая плотность в нижней части графика соответствует часто используемым словам, а верхняя кромка, в соответствии с законом Хипса [1], аппроксимируется функцией вида $u = n^{б}$, где $0 < б < 1$. Определенный интерес представляет «ширина» верхней кромки, которая соответствует повышенной частоте встречаемости слов после их первого появления (например, в диапазоне 250 слов исходного текста после первого появления заданного уникального слова). Для

информационного потока, состоящего из политематических документов, данный эффект объясняется различной лексикой, соответствующей различным тематикам, а для художественных произведений, по-видимому, – психологическими особенностями, механизмами извлечения слов авторами из своей памяти. На рис. 2 приведен график рангового распределения количества новых слов в заданном окне наблюдения для потока Интернет-новостей, сканируемых системой InfoStream [2], удовлетворительно аппроксимирующееся степенной функцией, что позволило сделать предположение о фрактальной природе рассматриваемого ряда [3].

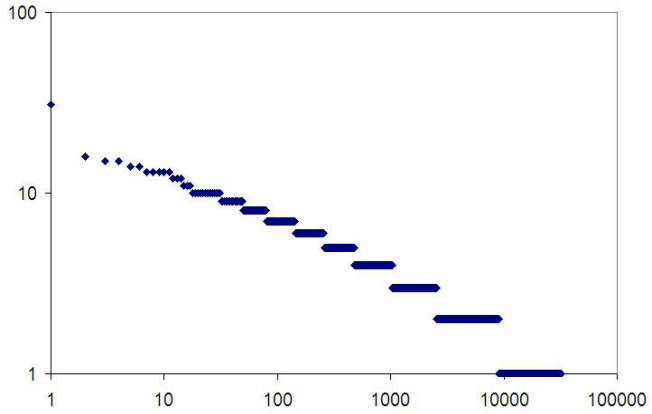

Рис. 2. Ранжированное распределение количества новых слов в окне наблюдения шириной в 250 слов

Полученная в соответствии с методикой, описанной в [4] соотношение нормированного размаха к стандартном отклонению значений этого ряда *R/S* (рис. 3) подтверждает высокую степень самоподобия рассматриваемого ряда (показатель Херста для различных текстовых массивов составляет от 0,65 до 0,8).

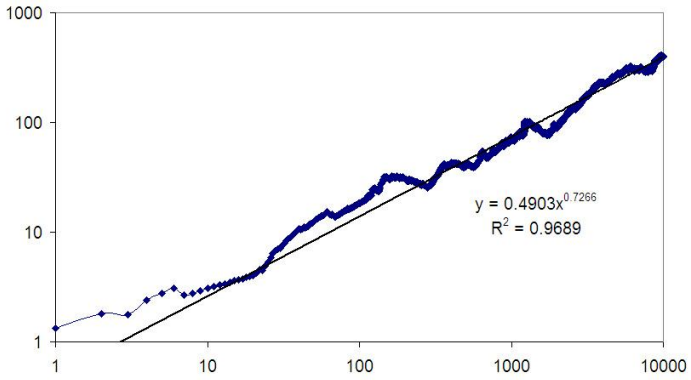

Рис. 3. Зависимость R/S для распределения количества новых слов в окне наблюдения шириной в 250 слов (в логарифмической шкале)

Авторам кажется перспективным усовершенствование метода анализа средней частоты встречаемости новых слов в окне наблюдений для автоматического определения жанра текстовых документов. Пока получены первые предварительные

результаты: при окне наблюдений в 250 слов эта величина для романа Л.Толстого «Анна Каренина» составила 1,1, в то время, как для политематического потока из Интернет – в среднем 1,25. Одновременно, показатель Херста для первого случая составил 0,75, а для второго – в среднем 0,70.

Кроме того, предложенный поход позволяет: отслеживать появление новой лексики; анализировать развитие тематических сюжетов в информационном потоке; обнаруживать перепечатки, информационные дубликаты, характеризующиеся отсутствием уникальных для документального потока лексических единиц.

ЛИТЕРАТУРА

1. Heaps H.S. Information Retrieval – Computational and Theoretical Aspects. Academic Press, 1978. – 344 p.
2. Григорьев А.Н., Ландэ Д.В., Бороденков С.А., Мазуркевич Р.В., Пацьора В.Н. InfoStream. Мониторинг новостей из Интернет: технология, система, сервис: научно-методическое пособие – Киев: ООО "Старт-98", 2007. – 40 с.
3. Федер Е. Фракталы. – М.: Мир, 1991. – 254 с.
4. Ландэ Д.В. Фрактальные свойства тематических информационных потоков из Интернет // Регистрация, хранение и обраб. данных, 2006. – Т. 8, № 2. – С. 93- 99.