一种表征字符信息量的法则：Character's Law*

李国华，咎红英
郑州大学 信息工程学院，郑州 450001
E-mail: liguohuahao@126.com; iehyzan@zzu.edu.cn

摘　要：Zipf 定律用于研究文本中单词出现频率与单词在频率表中的排名之间的关系，它在文献标引和词表编制、信息检索及图书情报管理中都有广泛应用。Heaps 定律研究文本中单词数目与文本大小的关系；Benford 定律研究现实生活数字中的首字母中 1~9 出现的规律。本文通过对英文基础词汇表观察和统计，发现一种表征字符信息量的法则：character's Law 即字符蕴含的信息量与字符频率排名的倒数符合 $F(x)=ax^b+c$ 分布，并对较大语料测试得到语料中同样符合幂分布。

关键词：Zipf's Law; Heaps' Law; Benford's Law; Character's Law; 字符信息量

Character's Law: Denotes the Value of the Character

Li Guohua, Zan Hongying
College of Information Engineering, Zhengzhou University, Zhengzhou 450001
E-mail: liguohuahao@126.com; iehyzan@zzu.edu.cn

Abstract: Zipf's Law is used to describe the relation between word's rank and word's frequency in a text. Zipf's Law can be widely applied to document indexing, vocabulary establish, information retrieval and information management. Heaps' Law described the relation between the size of word list and text length; Benford's Law states that the first digit is distributed in a specific, non-uniform way from many real-life sources of data. The paper proposed a character's law that the relation between a character's value and the reciprocal of its rank in frequency list is obey the distribution: $F(x) = ax^b+c$, and the distribution is same applies to the larger corpus.

Keywords: Zipf's Law; Heaps' Law; Benford's Law; Character's Law; character's value

1　前言

Zipf's Law(齐普夫定律)又称为最省力法则，它表述为：在自然语言书写的文本中，一个单词出现的频率与它在频率表中的排名成反比。Zipf定律由美国哈佛大学语言学家和心理学家George K. Zipf 1949年发表[1]，Zipf定律在文献标引和词表编制、信息检索、图书情报管理中都有应用。Wentian Li[2]利用随机生成的文本从某种程度上分析 Zipf 定律的有效性，文本中的每一个字符按照均匀分布生成，实验得出单词频率的分布和 Zipf 定律很接近。游荣彦[3]提出以 Zipf 定律仅描述汉字字频分布的尾部方法解决 Zipf 定律对整个汉字字频分布进行拟合调参时产生的误差。张化瑞[4]通过在不同大小的汉字集合上实验，得到汉字笔画数分布属于幂分布。

Heaps Law[5]研究给定文本中不同单词的数目，它指出不同单词的数目随文本大小的变化呈次线性增长，并与文本大小的平方根成正比[6]。Benford's Law[7]也称为首位数字定律(the first-digit law)，它由 1938 年物理学家 Frank Benford 提出：现实生活中数字的首字母中 1~9 出现的概率符合对数分布，“1”出现的概率为 30.1%，“2”出现的概率为 17.6%，“9”出现的概率只有 4.6%。

本文对英文基础词汇表中各个字符出现的规律进行了观察和统计，并定义字符表征单词的词汇数目和字符蕴含的信息量，得到字符的排名倒数与字符蕴含的信息量符合幂分布 $F(x)=ax^b+c$，实验得出在较大词汇表和较大语料中具有类似分布。

* 本文研究工作受到国家自然科学基金项目（项目号 60970083）和国家社会科学基金项目(项目号 09BTQ027)的资助。

2 Character's Law

我们定义词典中单词集为 W，|W|为词典中单词数目；单词 $w_i \in W$；n 为字符频率升序排名。字符 c 表征单词的信息量、频率和字符 c 表征的单词数及其比例的计算公式为(1)-(4)：

字符 c 的信息量定义为：$$weight(c) = \frac{\log(2 + numWord(c))}{\log(2 + frequency(c))} \tag{1}$$

字符 c 的频率定义为：$$frequency(c) = \sum_{w_i \in W} \begin{cases} 1 & c \in w_i \\ 0 & c \notin w_i \end{cases} \tag{2}$$

字符 c 表征的单词数为：$$numWord(c) = \sum_{w_i \in W} \begin{cases} 1 & \arg\min_{c_j \in w_i}(frequency(c_j)) = c \\ 0 & else \end{cases} \tag{3}$$

字符 c 表征单词数所占比例：$$percentage(c) = \frac{numWord(c)}{|W|} \tag{4}$$

3 实验及结果分析

对于字符频率的出现规律，本文使用两种方法进行验证：1）选择包含常用英语单词的词典 pocket 统计字符特征规律；2）选择词典 unixdict、词典 web2 和纯英文文本 bible、etext99 测试字符特征规律的有效性。

表 1　词典 Pocket 中字符特征

字符	排名	frequency	numWord	percentage	weight
j	1	282	282	0.013359	1.000000
q	2	352	350	0.016580	0.999035
z	3	430	419	0.019848	0.995750
x	4	447	436	0.020654	0.995938
k	5	1489	1426	0.067551	0.994092
w	6	1689	1446	0.068498	0.979129
v	7	1752	1576	0.074657	0.985844
f	8	2153	1766	0.083657	0.974211
y	9	2594	1738	0.082331	0.949110
b	10	3116	1967	0.093179	0.942863
g	11	3297	1831	0.086736	0.927461
h	12	3766	1498	0.070962	0.888141
m	13	4376	1851	0.087684	0.897453
p	14	4403	1566	0.074183	0.876892
d	15	4658	955	0.045239	0.812595
u	16	5504	607	0.028754	0.744384
c	17	6077	635	0.030081	0.741085
s	18	7137	354	0.016769	0.662089
l	19	7199	194	0.009190	0.594250
o	20	8638	83	0.003932	0.490134
n	21	8768	72	0.003411	0.474063
t	22	9346	41	0.001942	0.411379
i	23	9836	6	0.000284	0.226174
r	24	10192	9	0.000426	0.259806
a	25	10841	1	0.000047	0.118241
e	26	13167	1	0.000047	0.115819

3.1 统计 Character's Law

词典 pocket 包含 21110 个小写单词。根据表 1，本文得到在词典 pocket[1] 中字符的排名与字符表征单词数的关系(见图 1)；字符排名与字符的信息量之间的幂关系：$Y = ax^b + C$ (见图 2)，Y 为字符 c 的信息量的对数值；x 为字符 c 排名的倒数；拟合的参数值见表 2。

从图 2 中，可以得出，字符频率的排名越高，字符的信息量越大，即说明该字符越能表征单词。

表 2　图 2-10 中参数 a、b、c 的值

	a	b	c
图 2	-7.993e-005	-2.888	1
图 3	-6.896e-005	-2.917	1
图 4	-6.346e-009	-5.158	1.004
图 5	-2.827e-005	-3.221	0.9984
图 6	-2.926e-013	-7.32	1
图 7	-9.291e-016	-8.707	0.9975
图 8	-9.287e-016	-8.707	0.9975
图 9	-6.713e-016	-8.588	0.9881
图 10	-1.476e-015	-7.913	0.9893

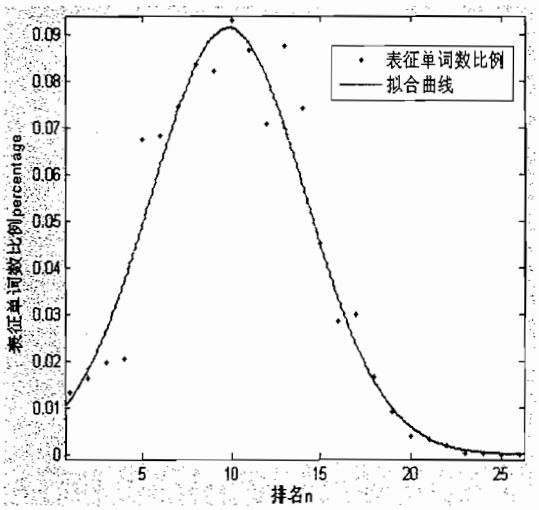

图 1　pocket 中字符排名 n 与表征单词数比例的关系

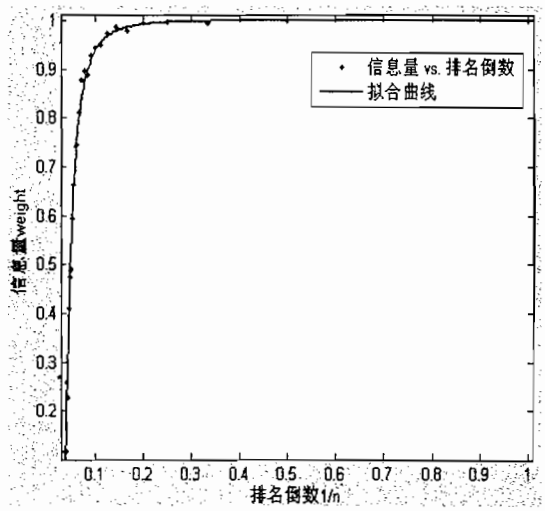

图 2　pocket 中字符排名倒数 $\frac{1}{n}$ 与信息量的关系

3.2 词典 unixdict[1] 和词典 web2[1]

词典 unixdict 出现于 Unix 大多数版本中，它使用小写字母形式并且不包含组合形式，包含 25104 个单词。词典 web2 由从 NI2（Webster's New International Dictionary, Second Edition，韦氏新国际词典第二版）中抽取的单个单词组成，包含 234936 个单词。图 3 和图 5 显示了在词典中字母依据出现频率的排名倒数 $\frac{1}{n}$ 与字符蕴含的信息量之间的关系；图 4 和图 6 显示了在词典中所有频率大于 0 的字符排名倒数 $\frac{1}{n}$ 与字符蕴含的信息量之间的关系。从图 3-4 中，可以发现二者分布均符合幂函数 $Y = ax^b + C$ 。

[1] http://www.puzzlers.org/dokuwiki/doku.php?id=solving:wordlists:about:start

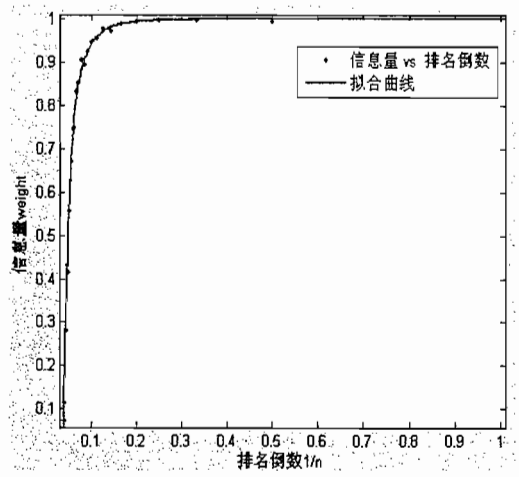

图3 unixdict 中字符排名倒数$\frac{1}{n}$与信息量的关系

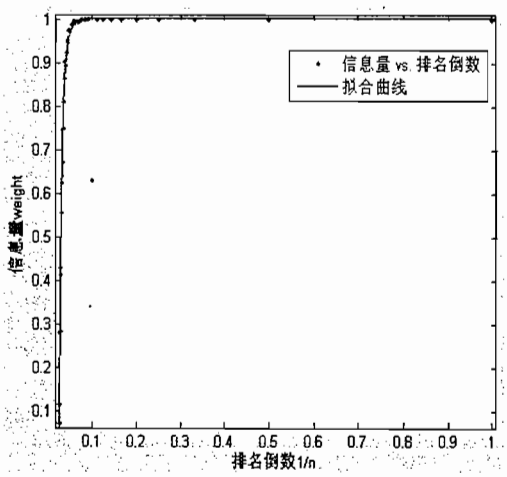

图4 unixdict 中字符排名倒数$\frac{1}{n}$与信息量的关系

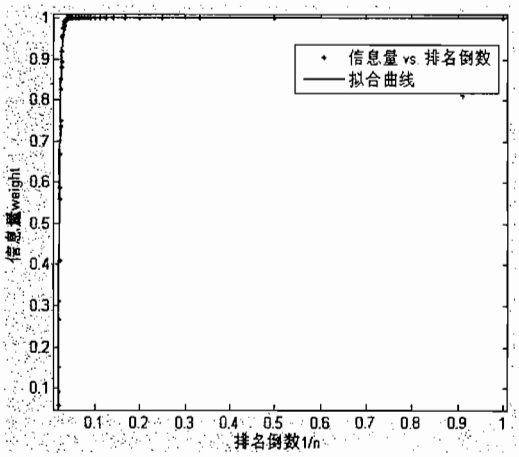

图5 web2 中字符排名倒数$\frac{1}{n}$与信息量的关系

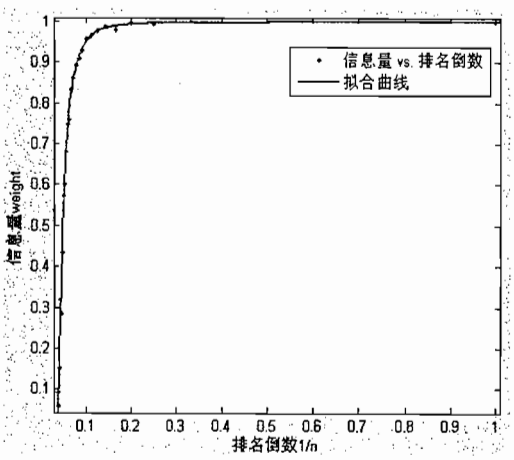

图6 web2 中字符排名倒数$\frac{1}{n}$与信息量的关系

3.3 文本 biblc[1]

文本 bible 大小为 3.85M，自然文本中，由于各个单词的数目不唯一，本文对自然文本中字符 c 的频率和表征的单词数同时乘以系数 count，count 为单词在文本中出现的次数。图 7 显示了在文本 bible 中字母排名的倒数$\frac{1}{n}$与字母蕴含信息量之间的关系；图 8 显示了再文本 bible 中字符排名的倒数$\frac{1}{n}$与字符蕴含的信息量之间的关系。

3.4 文本 etext99[2]

文本 etext99 大小为 100M，图 9 显示了在文本 etext99 中字母排名的倒数$\frac{1}{n}$与字母蕴含的信息

[1] http://corpus.canterbury.ac.nz/descriptions/#large

[2] http://people.unipmn.it/manzini/lightweight/corpus/

量之间的关系；图 10 显示了在文本 etext99 中频率大于 0 的字符的排名倒数$\frac{1}{n}$与字符蕴含的信息量之间的关系；从图 9~10 中，可以发现 Character's Law 在大数据集上的分布仍能够较好吻合；表 2 中显示出各个数据集上的差别仅于参数值的不同。

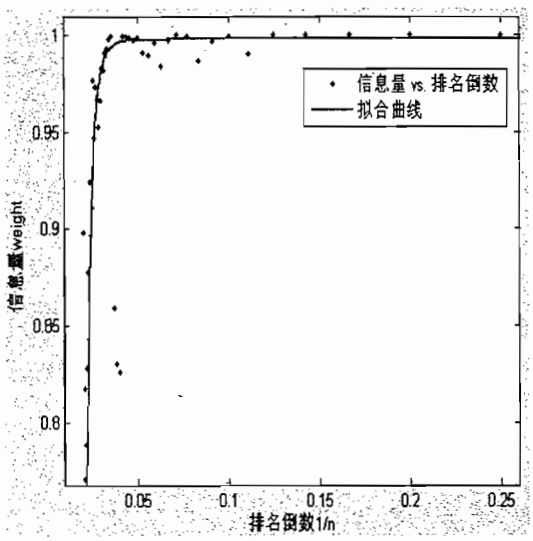

图 7　bible 中字符排名倒数$\frac{1}{n}$与信息量的关系

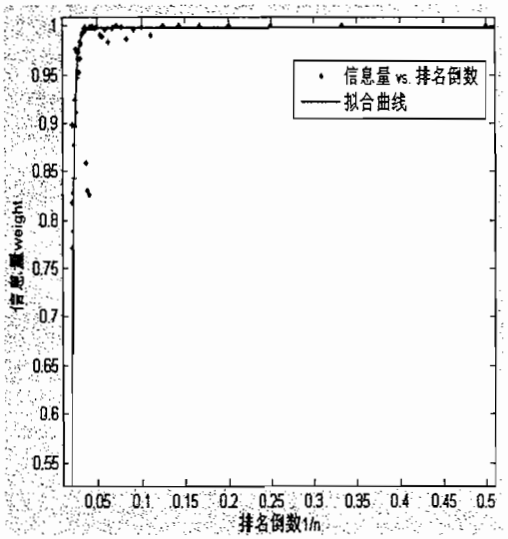

图 8　bible 中字符排名倒数$\frac{1}{n}$与信息量的关系

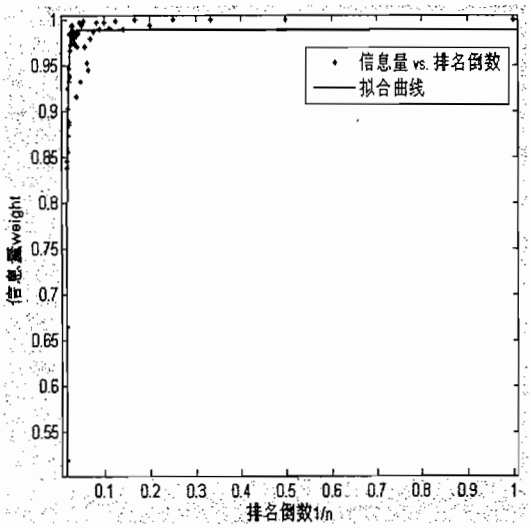

图 9　etext99 中字符排名倒数$\frac{1}{n}$与信息量的关系

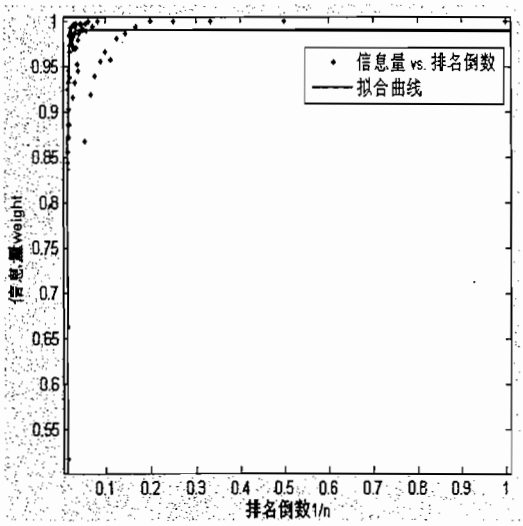

图 10　etext99 中字符排名倒数$\frac{1}{n}$与信息量的关系

4　结论与展望

本文通过对英文基础词汇表中各个字符出现的频率进行观察和统计，利用定义字符表征单词的词汇数目和字符蕴含的信息量，得到字符的排名倒数与字符蕴含的信息量符合幂分布 $F(x)=ax^b+c$，实验得出在较大词汇表和较大语料中具有类似分布，所不同的是不同语料，分布参数有差异。

本文预测，Character's Law 所表现的字符信息量会有更广泛的应用，下一步工作中，我们将研究 Character's Law 在大量查询字符串快速匹配中对数据集的预处理的作用以及在其他领域的潜在应用。

参 考 文 献

[1] George K. Zipf. Human Behaviour and the Principle of Least-Effort. Addison-Wesley, Cambridge MA, 1949.Dantchev S. Improved sorting-based procedure for integer programming. Mathematical Programming, Serial A, 2002, 92: 297-300.

[2] Wentian Li. Random Texts Exhibit Zipf's-Law-Like Word Frequency Distribution. IEEE Transactions on Information Theory, 1992, 38(6): 1842-1845.

[3] 游荣彦. Zipf 定律与汉字字频分布. 中文信息学报, 2000, 14(3): 60-65.

[4] 张化瑞. 汉字笔画数分布的一个统一公式. 第十一届汉语词汇语义学研讨会(CLSW2010), 2010: 477-482.

[5] Harold Stanley Heaps. Information Retrieval: Computational and Theoretical Aspects. Academic Press, 1978. Heaps' law is proposed in Section 7.5 (pp 206-208).

[6] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, Modern Information Retrieval, ACM Press, 1999.

[7] Frank Benford. The law of anomalous numbers. Proceedings of the American Philosophical Society, 1938, 78(4): 551-572.